AD_______________

Award Number: W81XWH-08-1-0110

TITLE: A Search for Gene Fusions/Translocations in Breast Cancer

PRINCIPAL INVESTIGATOR: Arul M. Chinnaiyan, M.D., Ph.D.

CONTRACTING ORGANIZATION: Regents of the University of Michigan
Ann Arbor, Michigan 48109-1274

REPORT DATE: October 2012

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE 21 October 2012	**2. REPORT TYPE** Annual	**3. DATES COVERED** 1 September 2011 – 31 August 2012

4. TITLE AND SUBTITLE

A Search for Gene Fusions/Translocations in Breast Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-08-1-0110

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Arul M. Chinnaiyan, M.D., Ph.D.

E-Mail: chinnaiyangrants@umich.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Regents of the University of
Michigan
Ann Arbor, Michigan 48109-1274

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Previously, we completed the molecular/ biochemical characterization of several shortlisted candidate gene fusions from the transcriptome sequencing of over 70 previously validated samples. From these studies, we identified two rare but recurrent gene fusions in breast cancer cell lines and tissues involving the MAST and Notch genes. Both of these fusion genes are potentially targetable and patients harboring MAST or Notch fusions may benefit from MAST or Notch inhibitors. Functional characterizations of the fusion genes in various in vivo and in vitro assays are ongoing. In this reporting period of our ongoing project, "Search for Gene Fusions/Translocations in Breast Cancer", we extended upon the analysis of gene fusions in breast cancer to examine amplicon-associated gene fusions and their functional significant. We also describe a novel study of cancer-specific pseudogenes.

15. SUBJECT TERMS
Gene fusions, sequencing, MAST,Notch

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	41 41	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

INTRODUCTION:

In our ongoing project entitled "A Search for Gene Fusions/Translocations in Breast Cancer" we have undertaken a systematic evaluation of breast cancer to map disease-specific, recurrent chromosomal or transcriptional chimeras in breast cancer that can be further characterized to develop novel biomarkers and therapeutic targets. Earlier, we reported the characterization of a subset of ER positive breast cancer patients characterized by the overexpression of AGTR1 who may be responsive to an available drug, losartan1 (Rhodes et al, 2009). We also provided a novel mechanistic framework for the overexpression of the polycomb group protein EZH2 in metastatic breast and prostate cancers, involving the genomic loss of its negative regulator, miR101 (Varambally et al, 2008). Additionally, we had reported high throughput sequencing pipeline for a directed search for gene fusions in cancers using next generation transcriptome sequencing platforms (Maher et al, 2009). From those efforts, we had identified numerous gene fusions (70 in over 40 cancer samples) that mapped to *loci* of genomic amplifications. We shortlisted several fusion candidates that involved kinase genes and other genes of interest related to oncogenesis for further study.

Previously we described the exciting discovery and characterization of two novel recurrent and actionable gene fusions in our breast cancer cohort involving MAST and Notch genes. Both MAST and Notch family gene fusions exerted significant phenotypic effects in breast epithelial cells (Robinson et al, 2011). We also reported the development of a novel bioinformatics tool designed to facilitate the discovery of gene fusions from next-generation sequencing data (Iyer et al, 2011); as well as a study that furthers our understanding of the role of microRNAs in cancer progression (Cao et al, 2011).

In this reporting period, we extended upon the analysis of gene fusions in breast cancer and a novel study of cancer-specific pseudogenes.

BODY:

A detailed, itemized report of the progress in work follows:

1. Characterization of recurrent gene fusions in breast cancer:

   Application of high-throughput transcriptome sequencing has spurred highly sensitive detection and discovery of gene fusions in cancer, but distinguishing potentially oncogenic fusions from random, "passenger" aberrations has proven challenging. We examined a distinctive group of gene fusions that involve genes present in the loci of chromosomal amplifications—a class of oncogenic aberrations that are widely prevalent in breast cancers. Integrative analysis of a panel of 14 breast cancer cell lines comparing gene fusions discovered by high-throughput transcriptome sequencing and genome-wide copy number aberrations assessed by array comparative genomic hybridization led to the identification of 77 gene fusions, of which more than 60% were localized to amplicons including 17q12, 17q23, 20q13, chr8q, among others. Many of these fusions appeared to be recurrent or involved highly expressed oncogenic drivers, frequently fused with multiple different partners, but sometimes displaying loss of functional domains.

Here we carried out a systematic analysis of the association between gene fusions and genomic amplification by integrating RNA-Seq data with array comparative genomic hybridization (aCGH)–based whole genome copy number profiling from a panel of breast cancer cell lines. We examined a set of “amplicon-associated gene fusions” that refer to all the fusions where one or both gene partners are localized to a site of chromosomal amplification. We found that as many as 62% of the total number of fusions were associated with regions of amplifications (Figure 1).

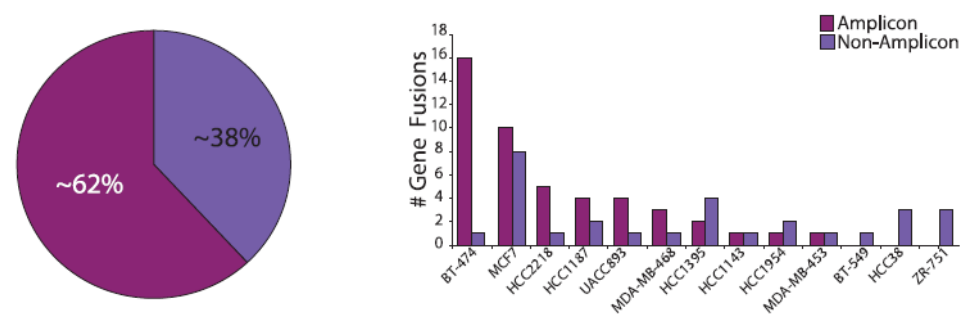


Figure 1. Distribution of gene fusions across breast cancer cell lines. Pie chart representation of the relative proportion of gene fusions associated with loci of genomic amplifications compared to unamplified loci (left) and bar graph representation of the relative distribution of gene fusions across different breast cancer cell lines (right). *Neoplasia*. 2012 Aug;14(8):702-8

We next assessed the functional relevance of two amplicon-associated fusion genes involving oncogenic kinases, EGFR and RPS6KB1, in the context of prioritizing fusion candidates important in tumorigenesis. In our transcriptome sequencing compendiumof 89 breast cancer cell lines and tissues, the highest expression of EGFR is observed inMDA-MB-468, potentially resulting from a focal amplification at chr7p12. In addition, we detected an EGFR fusion transcript (EGFR-POLD1) in this cell line, encoding the N-terminal portion of EGFR, completely devoid of the tyrosine kinase domain. Considering that the MDA-MB-468 harbors both MAST2 and EGFR fusions, we wanted to assess its relative “dependence” on both the kinases. Surprisingly, a profound reduction in cell proliferation was observed on siRNA knockdown of MAST2, whereas EGFR knockdown showed little effect (Figure 2). Next, testing the possibility of EGFR amplicon potentially cooperating with MAST2, we found that the effect of

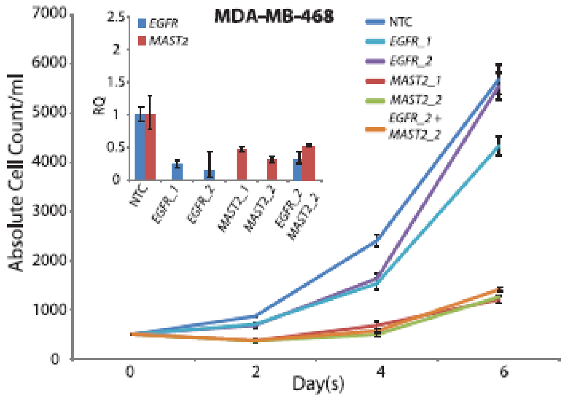


Figure 2. Proliferation assay showing absolute cell count (y axis) over a time course (x axis) after knockdown with EGFR and/or MAST2 siRNAs in MDA-MB-468. QPCR assessment of knockdown efficiencies relative to nontargeted control (NTC; inset). *Neoplasia*. 2012 Aug;14(8):702-8

combined knockdown of EGFR and MAST2 was comparable with that of MAST2 knockdown alone (Figure 2), further suggesting that EGFR amplification does not signify a driver aberration.

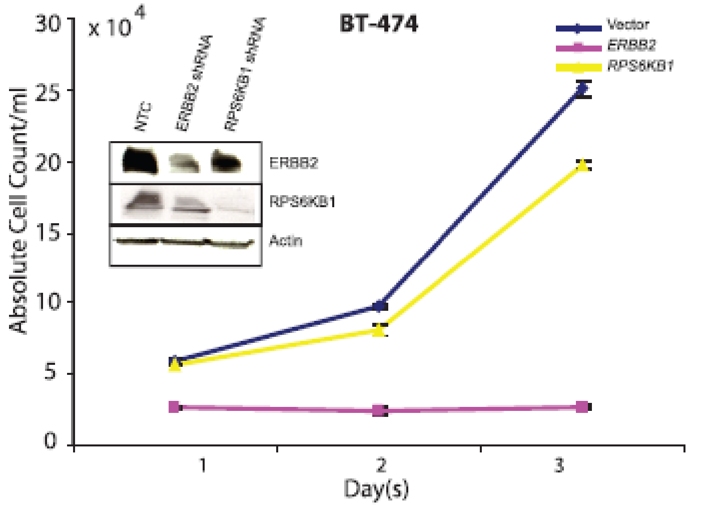


Figure 3. Proliferation assay showing absolute cell count (y axis) over a time course (x axis) after knockdown with EGFR and/or MAST2 siRNAs in MDA-MB-468. QPCR assessment of knockdown efficiencies relative to nontargeted control (NTC; inset). *Neoplasia*. 2012 Aug;14(8):702-8

Next, considering that BT-474 is an ERBB2-positive cell line, we tested potential dependence of these cells on the RPS6KB1 protein. Surprisingly, similar to our observations with EGFR knockdown in MDA-MB-468 cells, here we observed only a small effect on cell proliferation after shRNA knockdown of RPS6KB1, in dramatic contrast to the effect of ERBB2 knockdown (Figure 3). Notably, the shRNA knockdown of RPS6KB1 led to a significant depletion of the full-length protein yet it did not affect cell proliferation compared with ERBB2 protein depletion (Figure 3, inset). Therefore, BT-474 cells do not display a dependence on RPS6KB1 protein, and considering that the RPS6KB1 fusion product is completely devoid of all functional domains of RPS6KB1, including the kinase domain, this fusion also likely represents a passenger event.

Overall, our study suggests that amplicon-associated gene fusions in breast cancer primarily represent a by-product of chromosomal amplifications that constitutes a subset of passenger aberrations and should be factored accordingly during prioritization of gene fusion candidates. (Neoplasia. 2012 Aug;14(8):702-8).

2. Next Generation Sequencing Analysis:

Pseudogenes are a class of non-coding RNA transcripts that are dysfunctional relatives of known functional genes that have lost their protein coding ability and often not expressed. Aberrant expression of several functional non-coding RNA in cancer has been previously described, however genome-wide expression of pseudogenes had not been reported for any cancer type. We developed a pseudogene expression pipeline to analyze a large compendium of paired-end next generation sequencing (RNASeq) data generated from 293 samples, comprising 13 different epithelial cancers. Our integrative approach provided evidence of expression for 2,082 distinct pseudogenes that displayed lineage-specific, cancer-specific, as well as ubiquitous expression patterns.

Though a majority of the pseudogenes examined were found in both cancer and benign samples, we observed 218 pseudogenes expressed only in cancer samples, of which 178 were observed in multiple cancers and 40 were found to have highly specific expression in a single cancer type only (Figure 4).

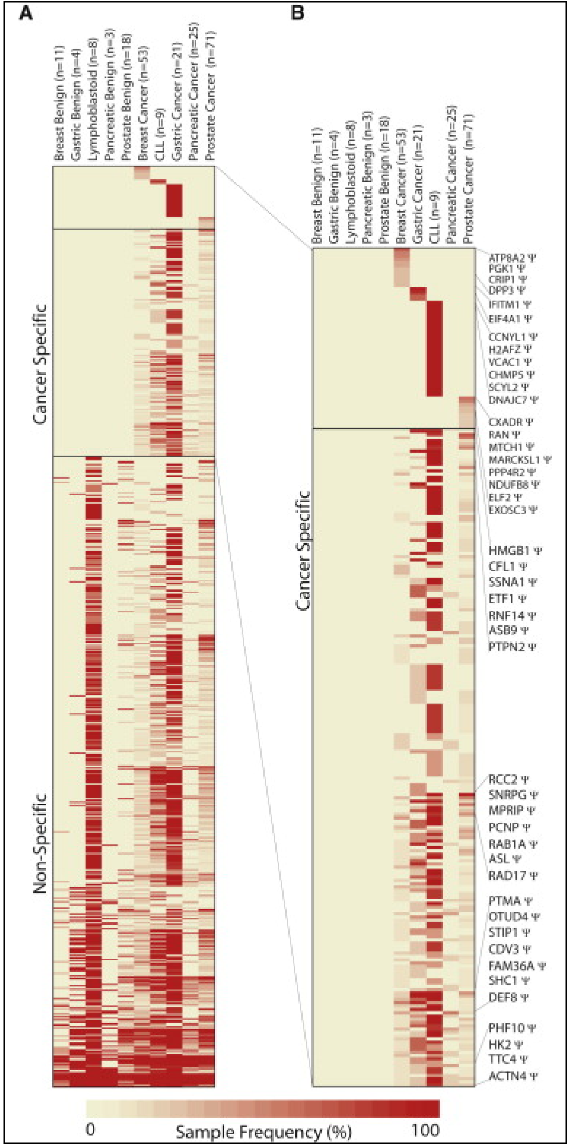


Figure 4. Cancer-Specific Pseudogene Expression Profiles(A) Heatmap of pseudogene expression sorted according to cancer-specific expression patterns displays pseudogene transcripts specific to individual cancers (top), common across multiple cancers (tissue-enriched; middle), and nonspecific (bottom).(B) Zoomed-in version of the top panel displaying individual cancer-specific expressed pseudogenes. The columns represent different tissues with the number of samples in parentheses. The rows represent individual clusters mapping to specific pseudogenes. The color intensity represents the frequency (%) of samples in a tissue type showing expression of a given pseudogenes (according to the scale indicated at the bottom). *Cell*. 2012 Jun 22;149(7):1622-34.

Among the pseudogene candidates in breast cancer, we identified an unprocessed pseudogene cognate to ATP8A2, a LIM domain-containing protein speculated to be associated with stress response and proliferative activity. ATP8A2-Ψ expression found to be restricted to breast samples, the highest levels seen in a subset of breast cancer tissues and cell lines (Figure 4). By contrast, ATP8A2-WT expression was highly variable across different tissue types and showed no correlation with ATP8A2-Ψ expression. To investigate a potential role of ATP8A2-Ψ expression in breast cancer, first we carried out siRNA-based knockdown of both the wild-type and pseudogene RNA in two independent breast cancer cell lines that expressed both the transcripts. Knockdown of ATP8A2-Ψ with two independent siRNAs was found to specifically inhibit the proliferation of overexpressing cell lines Cama-1 and HCC1806 (Figure 5A), but not the cell lines with no detectable levels of ATP8A2-Ψ, for example, the benign breast epithelial cell line H16N2 (Figure 5A, right). Knockdown of ATP8A2-Ψ (but not ATP8A2-WT) also resulted in reduced cell migration and invasion seen in *in vitro* Boyden Chamber assays (Figure 5B) as well as in *in vivo* intravasation and metastasis in chicken chorioallantoic membrane xenograft assay (Figure 5C). In contrast, knockdown of wild-type ATP8A2 had no effect on the proliferation of any of the cell lines tested, suggesting an unexpected growth regulatory role for ATP8A2-Ψ.

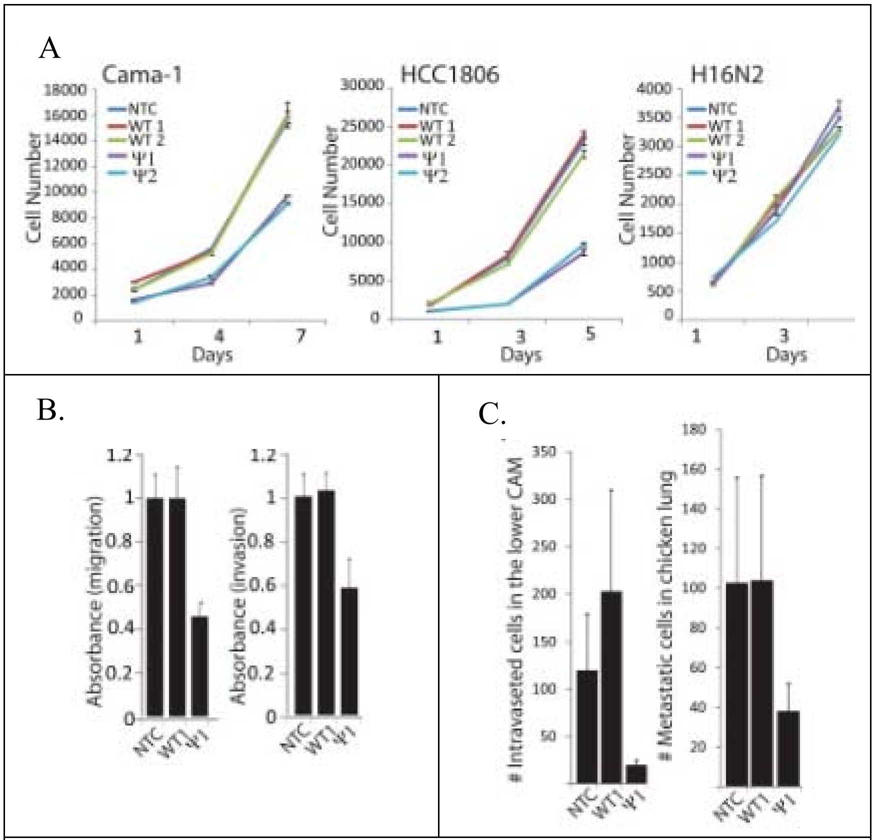


Figure 5. (A) Cell proliferation assays following siRNA knockdowns of ATP8A2-WT and -Ψ as indicated. NTC, nontargeting control; WT, siRNA against wild-type ATP8A2; Ψ, siRNA against ATP8A2-Ψ.(B) Boyden chamber assay showing cell migration (left) and invasion through matrigel (right). (C) chicken chorioallantoic membrane assay of HCC-1806 cells treated with nontargeting control siRNA, ATP8A2-WT, or ATPA2-Ψ siRNA showing relative number of cells intravasated in the lower CAM (left) and metastatic cells in chicken lung (right).Error bars represent means ± SE of the mean. *Cell.* 2012 Jun 22;149(7):1622-34.

This study is the first large-scale analysis of pseudogene expression in human cancer using transcriptome sequencing data. (Cell. 2012 Jun 22;149(7):1622-34).

KEY RESEARCH ACCOMPLISHMENTS:

- We performed an integrated analysis combining RNASeq and aCGH to examine amplicon-associated gene fusions across 14 breast cancer cell lines. We found that many of these fusions, even when they involve known oncogenes, are often “passenger” events that do not display oncogenic potential.
- We used a novel bioinformatics approach to analyze next generation sequencing data to discover novel expressed pseudogenes. Although many of the pseudogenes are ubiquitously expressed, we found a sub-set of them are expressed in a lineage and cancer-specific manner, including the breast cancer-specific pseudogene, ATP8A2Ψ.

REPORTABLE OUTCOMES:

Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu YM, Cao X, Asangani IA, Kothari V, Prensner JR, Lonigro RJ, Iyer MK, Barrette T, Shanmugam A, Dhanasekaran SM, Palanisamy N, Chinnaiyan AM. Expressed pseudogenes in the transcriptional landscape of human cancers. Cell. 2012 Jun 22;149(7):1622-34.
PubMed PMID: 22726445.

Kalyana-Sundaram S, Shankar S, Deroo S, Iyer MK, Palanisamy N, Chinnaiyan AM, Kumar-Sinha C. Gene fusions associated with recurrent amplicons represent a class of passenger aberrations in breast cancer. Neoplasia. 2012 Aug;14(8):702-8. PubMed PMID: 22952423; PubMed Central PMCID: PMC3431177.

CONCLUSION:

This past year we extended upon our previous studies that identified two rare but recurrent gene fusions in breast cancer cell lines and tissues involving the potentially actionable MAST and Notch genes. We analyzed amplicon-associated gene fusions across 14 cell lines and conclude that many of them are likely passenger events that are not oncogenic drivers. In addition we used a novel bioinformatics approach to discover expressed pseudogenes. Of particular interest are those that display cancer-specific expression, including breast cancer, and confer cell proliferative and metastatic properties. This represents another layer of complexity of cancer biology that was previously unappreciated.
"So What?: The tools that we develop to identify novel gene fusions and other drivers of tumorigenesis along with the biological functional analysis lays the framework for developing personalized breast cancer therapies based on driving mutation.

REFERENCES:

1. Varambally S, Cao Q, Mani RS, Shankar S, Wang X, Ateeq B, Laxman B, Cao X, Jing X, Ramnarayanan K, Brenner JC, Yu J, Kim JH, Han B, Tan P, Kumar-Sinha C, Lonigro RJ, Palanisamy N, Maher CA, Chinnaiyan AM: Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer, Science 2008, 322:1695-1699.

2. Rhodes DR, Ateeq B, Cao Q, Tomlins SA, Mehra R, Laxman B, Kalyana-Sundaram S, Lonigro RJ, Helgeson BE, Bhojani MS, Rehemtulla A, Kleer CG, Hayes DF, Lucas PC, Varambally S, Chinnaiyan AM: AGTR1 overexpression defines a subset of breast cancer and confers sensitivity to losartan, an AGTR1 antagonist, Proc Natl Acad Sci U S A 2009, 106:10284-10289.

3. Ateeq B, Tomlins SA, Chinnaiyan AM. AGTR1 as a therapeutic target in ER-positive and ERBB2-negative breast cancer cases. Cell Cycle. 2009 Dec;8(23):3794-5. PubMed PMID: 19934656; PubMed Central PMCID: PMC2940713.

4. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette TR, Grasso C, Yu J, Lonigro RJ, Schroth G, Kumar-Sinha C, Chinnaiyan AM. Chimeric transcript discovery by paired-end transcriptome sequencing, Proc Natl Acad Sci U S A 2009.

5. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM. Transcriptome sequencing to detect gene fusions in cancer. Nature 2009;458(7234):97-101.

6. Palanisamy N, Ateeq B, Kalyana-Sundaram S, Pflueger D, Ramnarayanan K, Shankar S, Han B, Cao Q, Cao X, Suleman K, Kumar-Sinha C, Dhanasekaran SM, Chen YB, Esgueva R, Banerjee S, LaFargue CJ, Siddiqui J, Demichelis F, Moeller P, Bismar TA, Kuefer R, Fullen DR, Johnson TM, Greenson JK, Giordano TJ, Tan P, Tomlins SA, Varambally S, Rubin MA, Maher CA, Chinnaiyan AM: Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma, Nat Med 2010; 16:793-798.

7. Cao Q, Mani RS, Ateeq B, Dhanasekaran SM, Asangani IA, Prensner JR, Kim JH, Brenner JC, Jing X, Cao X, Wang R, Li Y, Dahiya A, Wang L, Pandhi M, Lonigro RJ, Wu YM, Tomlins SA, Palanisamy N, Qin Z, Yu J, Maher CA, Varambally S, Chinnaiyan AM. Coordinated regulation of polycomb group complexes through microRNAs in cancer. Cancer Cell. 2011 Aug 16;20(2):187-99. PubMed PMID: 21840484; PubMed Central PMCID: PMC3157014.

8. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011 Oct 15;27(20):2903-4. Epub 2011 Aug 11. PubMed PMID: 21840877; PubMed Central PMCID: PMC3187648.

9. Robinson D.R., Kalyana-Sundaram S., Wu Y.-I., Shankar S., Cao X., Ateeq B., Asangani I.A., Iyer M., Maher C.A., Grasso C.S., Lonigro R.J., Quist M., Siddiqui J., Mehra R., Jing X., Giordano T.J., Sabel M.S., Kleer C.G., Palanisamy N., Natrajan R., Lambros M.B., Reis-Filho J.S., Kumar-Sinha C., and Chinnaiyan A.M. Functionally Recurrent Rearrangements of the MAST Kinase and Notch Gene Families in Breast Cancer. Nat Med. 2011 Nov 20;17(12):1646-51. PMID: 22101766; PMCID: PMC3233654.

UPDATED EXISTING/PENDING SUPPORT STATEMENT (UEPS)

EFFORTS IN BREAST CANCER RESEARCH: 50% Total

ACTIVE

W81XWH-08-0110 (PI: Chinnaiyan)	09/01/08 – 08/31/13	3.0 cal mos
Department of Defense – Era of Hope	$500,000/yr	

A Search for Gene Fusions/Translocations in Breast Cancer

Specific Aims: 1) develop high-throughput adaptations of existing methodologies such as fluorescence in situ hybridization (FISH), 2) employ bioinformatics and associated analytical tools to elucidate recurrent gene fusions in breast cancers, 3) employ next generation whole transcriptome sequencing of breast tumors.

Contact Information at funding agency: Grants Officer: Cheryl A. Lowery, 301-619-7150, Cheryl.Lowery@us.army.mil, U.S. Army Medical Research Acquisition Activity, 820 Chandler Street (MCMR-AAA-R), Fort Detrick, MD 21702-5014

Effort to breast cancer: 25%

W81XWH-12-1-0080 (PI: Chinnaiyan)	09/15/12 – 09/14/17	1.2 cal mos
Department of Defense	$479,470/yr	

Advancing our understanding of the etiologies and mutational landscapes of basal-like, luminal A, and luminal B breast cancers

Specific Aims: 1) Identify and quantify risk factors for each of the most common molecular subtypes of breast cancer, basal-like, luminal A, and luminal B tumors, in a large-scale population-based study. 2) Discover and validate the mutational landscape of basal-like, luminal A, and luminal B tumors. 3) Characterize the relationships between subtype specific risk factors and mutational signatures. 4) Develop and validate risk prediction models unique to each breast cancer subtype incorporating clinical, epidemiologic and mutation data. 5) Identify and quantify the relationships between various exposures and mutational changes on risk of breast cancer recurrence and survival among patients with basal-like, luminal A, and luminal B tumors.

Contact Information at funding agency: Cheryl A. Lowery, U.S. Army Medical Research Acquisition Activity, 820 Chandler Street (MCMR-AAA-R), Fort Detrick, MD 21702-5014, 301-619-7150, Cheryl.Lowery@us.army.mil

Effort to breast cancer: 10%

PI: Chinnaiyan	01/01/09 – 12/31/13	1.2 cal mos
Doris Duke Foundation	$275,000/yr	

Distinguished Clinical Scientist Award for Excellence in "Bench to Bedside" Research

Goal(s): to launch a new effort in the laboratory to comprehensively and systematically scour common human solid tumors for the presence of recurrent gene rearrangements. This effort primarily funds the training of new translational researchers under the mentorship of Dr. Chinnaiyan.

Specific Aims: 1) Develop and employ high-throughput fluorescence in situ hybridization (FISH) in order to interrogate solid tumors for recurrent chromosomal aberrations including gene fusions and translocations; 2) Employ bioinformatics and associated analytical tools to elucidate recurrent gene fusions in common solid tumors;. 3) Employ next generation whole transcriptome and paired-end sequencing of common solid tumors to identify recurrent gene fusions and integrated non-human sequences that may represent pathogens.

Contact Information at funding agency: Grants Officer: Betsy Myers, emyers@ddcf.org, Doris Duke Charitable Foundation, 650 5th Avenue, Fl 19, NY, NY; Phone: 212-974-7000;

R01-HG-005119 (PI: Qin)	07/22/09 – 06/30/13	0.12 cal mos
NIH/NHGRI	$32,154/yr	

Model-Based Methods of Analyzing ChIP Sequencing Data

Goal(s): to demonstrate that effective data integration under a coherent probability framework will lead to an in-depth understanding of mechanisms mediating transcription regulation in cancer progression.

Role: Co-Investigator

Contact Information at funding agency: Teresa Sussman, Emory University, 1518 Clifton Road NE, 8th Fl, Atlanta, GA 30322, tpoint@emory.edu; Phone 404-727-2503

Effort to breast cancer: 10%

Summary of results: Tangible progress has been made in the development of new bioinformatics tools for recurrent gene fusion discovery in solid tumors. Our COPA (Cancer Outlier Profile Analysis) approach to gene fusion discovery was detailed in our initial proposal, and we have since expanded upon this concept with applications to breast cancer. We sought to identify driving gene fusions in breast cancer using a combination of a meta-analysis for gene expression and validation across 31 gene expression studies and approximately 3200 microarray experiments. The culmination of these efforts were reported with the discovery that the gene AGTR1 (angiotensin II receptor type 1) was over expressed in 10-20% of the 311 breast cancer cases tested by FISH (Rhodes, et al, 2009). The highest overexpression of AGTR1 occurred in estrogen responsive ERBB2-negative tumors. This is particularly exciting, in that AGTR1 has already been targeted by an existing hypertension medication (Losartan) which has already been FDA approved. *In vivo* studies performed in a mouse model show promising results for the possible use of this agent in treating breast cancer patients with AGTR1-overexpressing breast tumors.

A new, sensitive, high-throughput analytic method was developed by our group to mine for functional gene fusions using paired-end transcriptome sequencing (Maher, et al., PNAS, 2009). We performed paired-end analysis on the MCF-7 breast cancer cell line, which resulted in the identification of an additional five new mutations in breast cancer. We identified two recurrent, actionable gene fusions in a subset of breast cancer cohorts, involving the MAST and Notch genes. Moreover, breast cancer patients harboring MAST or Notch fusions could potentially be treated with their respective inhibitors.

2U01CA111275-06 (PI: Chinnaiyan)	08/01/10 – 06/30/15	1.2 cal mos
NIH/NCI	$370,000/yr	

EDRN: A Systems Biology Approach to the Development of Cancer Biomarkers

Goal(s): Develop “Omics based approaches to study solid tumor for the purpose of developing biomarkers.

Specific aims: Platform technologies cover epigenetics, genomics, transcriptomics, proteomics and metabolomics.

Contact Information at funding agency: Wendy Briscoe, briscoew@mail.nih.gov, Phone: 301-496-3160, Fax: 301-496-8601, EPS- 6120 Executive Blvd, 243, Rockville, MD 20852

Effort to breast cancer: 5%

Summary of results: While this grant has been focused on prostate cancer, in general it is a biomarker development lab and half of my effort can be designated to the development of breast cancer biomarkers including AGTR in ER+, erbB2- patients.

AACR (Dream team leader: Chinnaiyan) 08/01/12 – 8/31/15 1.2 cal mos
Stand up to Cancer and Prostate Cancer Foundation Dream Team $440,021/yr
Precision Therapy of Advanced Prostate Cancer
Goal(s): The overall goal of this proposal is to catalyze the interaction of a multi-disciplinary team of investigators, with a track record of accomplishments in prostate cancer research, to work together on the challenging problem of metastatic castration resistant prostate cancer (CRPC).
Specific Aims: 1) Establish a multi-institutional infrastructure incorporating 5 leading prostate cancer clinical sites, 2 sequencing and computational analysis sites, linked with appropriate sample and data coordination; 2) Establish a prospective cohort of 500 patients (the "CRPC 500") utilizing the multi-institutional infrastructure to support the clinical use of integrative prostate cancer sequencing, analysis, and clinical trial decision making; 3) Conduct parallel, preclinical *in vivo* functional studies of resistance biomarkers and of SU2C-PCF sponsored combination therapies; 4) Identify molecular determinants of abiraterone sensitivity and acquired resistance in patients; 5) Conduct clinical trials of novel combinations targeting AR and/or the PTEN pathway, based on existing preclinical data and an understanding of resistance mechanisms; 6) Identify molecular determinants of sensitivity and acquired resistance to PARP inhibitors in patients.
Contact Information at funding agency: Frederic Biemar, frederic.biemar@aacr.org, (215) 446-7261

W81XWH-11-1-0337 (PI: Chinnaiyan) 09/30/11 – 09/29/15 1.2 cal mos
Department of Defense $145,145/yr
Prostate Cancer Research Program Idea Development Award, Established Investigator
Discovery of Novel Gene Elements Associated with Prostate Cancer Progression
Goal(s): determine if prostate cancer harbors numerous uncharacterized ncRNAs and show that a subset of these is differentially-expressed transcripts
Specific Aims: 1) to employ next generation sequencing to comprehensively annotate expressed regions in the prostate cancer transcriptome; 2) to validate and characterize transcriptional units in poorly annotated regions; 3) to elucidate a functional and clinical role of poorly-annotated transcripts in prostate cancer.
Contact Information at funding agency: Janet Kuhns, 301-619-2827, janet.kuhns@us.army.mil, U.S. Army Medical Research Acquisition Activity, 820 Chandler Street (MCMR-AAA-R), Fort Detrick, MD 21702-5014

R01CA154365 (PIs: Beer, Chinnaiyan) 12/01/10 – 11/30/15 0.36 cal mos
NIH/NCI $85,529/yr
Identification and Characterization of Gene Fusions in Lung Adenocarcinoma
Goal(s): are to identify recurrent gene fusions in human lung adenocarcinoma and to determine the functional consequences of their action in lung adenocarcinoma cell lines. Prioritization will be to examine those lung gene fusions that may be therapeutically targetable.
Contact Information at funding agency: Rebecca Brightful, **Email**: brightfr@mail.nih.gov **Phone**: 301-631-3011 **Fax**: 301-451-5391,

R01CA132874-01 (PI: Chinnaiyan) 03/01/09 – 12/31/13 0.96 cal mos
NIH/NCI $166,000/yr
Molecular Sub-typing of Prostate Cancer Based on Recurrent Gene Fusions
Goal(s): Identification of novel molecular subtypes of cancer, characterization of these subtypes, and correlation of these with disease outcome using prostate needle biopsy samples.

Specific Aims: 1) discovery and nomination of novel molecular sub-types of prostate cancer, 2) characterize associations of molecular sub-types of prostate cancer with clinical outcome and/or aggressiveness of disease in a radical prostatectomy cohort, 3) characterize associations of molecular sub-types of prostate cancer with clinical outcome and/or aggressiveness of disease using prostate needle biopsy samples.
Contact Information at funding agency: Grants Management Specialist: Rebecca Brightful, Email: brightfr@mail.nih.gov Phone: 301-631-3011

P50 CA69568 (PI: Pienta)	06/01/08 - 05/31/13	0.78 cal mos
NIH/NCI	$177,509/yr	

SPORE in Prostate Cancer
Project 1 Title: *Role of gene fusions in prostate cancer*
Goal(s): determine the role of ETS family gene fusions in prostate cancer cell lines; characterize the phenotype of androgen-regulated ETS transgenic mice.
Specific Aims: 1) Characterization of Oncogenic ETS Gene Fusions in Prostate Cancer; 2) Determine the role of ETS family gene fusions in prostate cancer cell lines; 3) characterize the phenotype of androgen-regulated ETS transgenic mice.
Role: Co-Investigator
Contact Information at funding agency: Andrew Hruszkewycz, 301-496-8528, hruszkea@mail.nih.gov

. P50 CA69568 (PI: Pienta)	06/01/08 – 05/31/13	0.48 cal mos
SPORE in Prostate Cancer	$306,062/yr	

Core 3: Tissue/Informatics Core Director NIH/NCI Goal(s): the goal of the Core is to collect biological material with associated clinical information to facilitate translational research.
Role: Core Director
Contact Information at funding agency: Andrew Hruszkewycz, 301-496-8528, hruszkea@mail.nih.gov

PENDING:

1UM1HG006508-01A1 (PI: Chinnaiyan)	03/01/13 – 02/28/17	1.2 cal mos
National Institutes of Health	1,500,000/yr	

Exploring Precision Cancer Medicine for Sarcoma and Rare Cancers
Specific Aims: Project 1) Clinical Genomic Study, 1) Accrue 500 patients with advanced or refractory rare cancer for participation in an integrated approach to Clinical Genomics; 2) Interpret results through a multi-disciplinary Sequencing Tumor Board and disclose results to patients and their physicians; 3) Measure the influence of sequence results provided to patients; 4) Determine the frequency of clinically significant germline mutations in patients undergoing comprehensive tumor sequence analysis.
Project 2) Sequencing, Analysis, and Interpretation of Sequencing Data; 1) Process and track specimens and ensure quality control; 2) Sequence tumor and germline biospecimens; 3) Analyze sequencing data to identify clinically significant variants; 4) Interpret and translate sequence variants into clinical oncology setting; 5) Assess and evaluate costs associated with clinical sequencing.

OVERLAP FOR ALL CURRENT AND PENDING GRANTS:

None.

Expressed Pseudogenes in the Transcriptional Landscape of Human Cancers

Shanker Kalyana-Sundaram,[1,2,6,7] Chandan Kumar-Sinha,[1,2,7] Sunita Shankar,[1,2] Dan R. Robinson,[1,2] Yi-Mi Wu,[1,2] Xuhong Cao,[1,3] Irfan A. Asangani,[1,2] Vishal Kothari,[1] John R. Prensner,[1,2] Robert J. Lonigro,[1,2] Matthew K. Iyer,[1] Terrence Barrette,[1,2] Achiraman Shanmugam,[6] Saravana M. Dhanasekaran,[1,2] Nallasivam Palanisamy,[1,2] and Arul M. Chinnaiyan[1,2,3,4,5,*]

[1]Michigan Center for Translational Pathology
[2]Department of Pathology
[3]Howard Hughes Medical Institute
[4]Department of Urology
[5]Comprehensive Cancer Center
University of Michigan, Ann Arbor, MI 48109, USA
[6]Department of Environmental Biotechnology, Bharathidasan University, Tiruchirappalli 620 024, India
[7]These authors contributed equally to this work
*Correspondence: arul@umich.edu

DOI 10.1016/j.cell.2012.04.041

SUMMARY

Pseudogene transcripts can provide a novel tier of gene regulation through generation of endogenous siRNAs or miRNA-binding sites. Characterization of pseudogene expression, however, has remained confined to anecdotal observations due to analytical challenges posed by the extremely close sequence similarity with their counterpart coding genes. Here, we describe a systematic analysis of pseudogene "transcription" from an RNA-Seq resource of 293 samples, representing 13 cancer and normal tissue types, and observe a surprisingly prevalent, genome-wide expression of pseudogenes that could be categorized as ubiquitously expressed or lineage and/or cancer specific. Further, we explore disease subtype specificity and functions of selected expressed pseudogenes. Taken together, we provide evidence that transcribed pseudogenes are a significant contributor to the transcriptional landscape of cells and are positioned to play significant roles in cellular differentiation and cancer progression, especially in light of the recently described ceRNA networks. Our work provides a transcriptome resource that enables high-throughput analyses of pseudogene expression.

INTRODUCTION

Pseudogenes are ancestral copies of protein-coding genes that arise from genomic duplication or retrotransposition of mRNA sequences into the genome followed by accumulation of deleterious mutations due to loss of selection pressure, degenerating eventually into so-called genetic fossils (Sasidharan and Gerstein, 2008). Pseudogenes pervade the genome, representing virtually every coding gene, and due to their extremely close sequence similarity with their cognate genes, complicate whole-genome sequencing and gene expression analyses. A growing body of evidence strongly suggests their potential roles in regulating cognate wild-type gene expression/function by serving as a source of endogenous siRNA (Tam et al., 2008; Watanabe et al., 2008), antisense transcripts (Zhou et al., 1992), competitive inhibitors of translation of wild-type transcripts (Kandouz et al., 2004), and perhaps dominant-negative peptides (Katoh and Katoh, 2003). Pseudogene transcription has also been shown to regulate cognate wild-type gene expression by sequestering miRNAs (Poliseno et al., 2010). The recently described competing endogenous RNA (ceRNA) networks comprising sets of coordinately expressed genes with shared miRNA response elements (MREs) provide an additional dimension of (post-) transcriptional regulation in which the role of pseudogenes might overlap with those of protein-coding genes (Salmena et al., 2011; Sumazin et al., 2011).

Previous genome-wide studies of pseudogenes focused on the identification of their chromosomal coordinates and annotations based on diverse computational approaches (Karro et al., 2007; Zhang and Gerstein, 2004), including PseudoPipe (Zhang et al., 2006), HAVANA (Solovyev et al., 2006), PseudoFinder (Lu and Haussler, 2006, ASHG, conference), and Retrofinder (Zheng and Gerstein, 2006). These individual pipelines were subsequently consolidated into an integrated consensus platform, ENCyclopedia Of DNA Elements (ENCODE), which now serves as the definitive database of manually curated and annotated pseudogenes as well as pseudogene transcripts (Zheng et al., 2007). By contrast, genome-wide analyses of pseudogene expression have been somewhat arbitrary, mainly relying upon

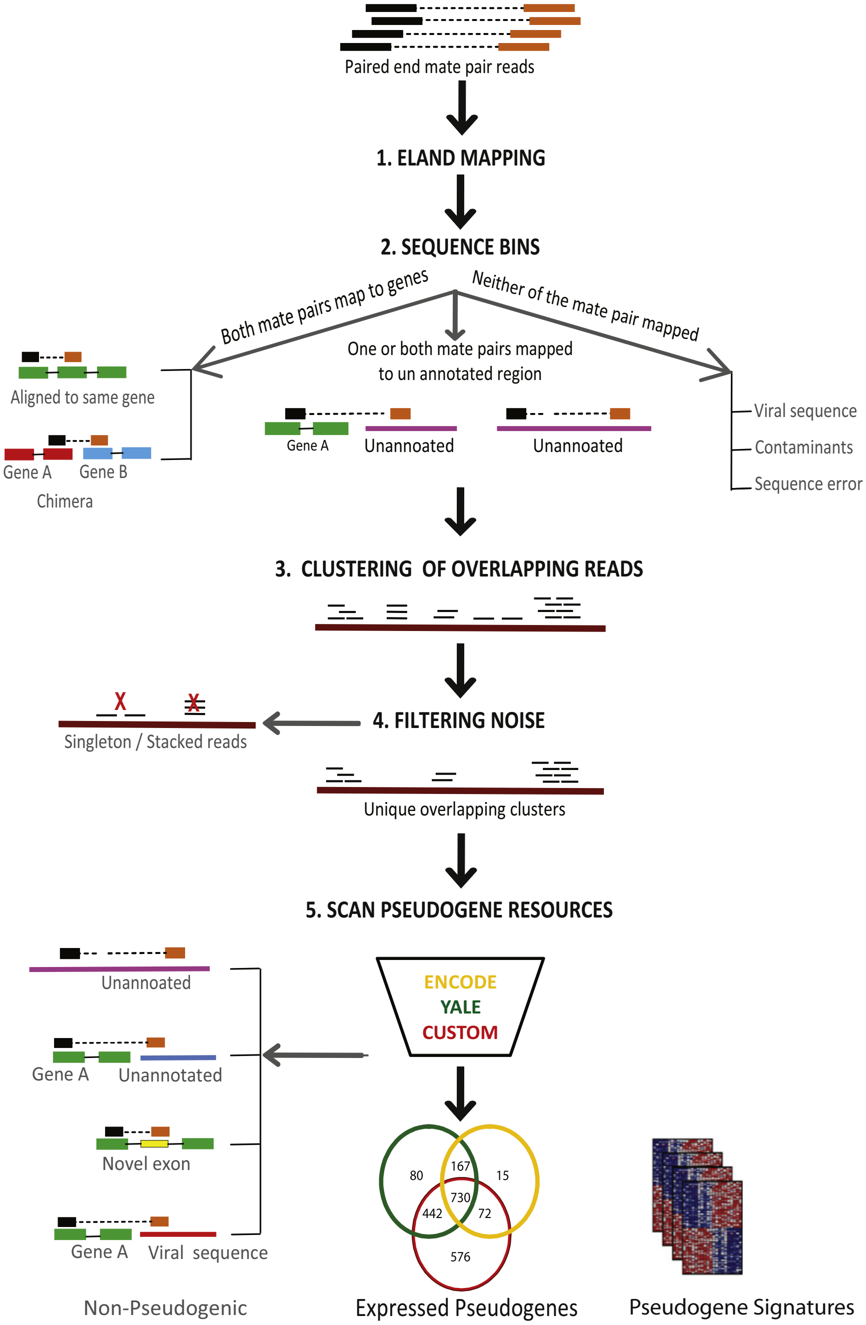


Figure 1. Pseudogene Expression Analysis Pipeline

The bioinformatics pipeline for analyzing pseudogene transcription involved the following steps: (1) Paired-end transcriptome sequencing reads were mapped to the human genome and UCSC Genes using ELAND. (2) Passed purity (PF) filter reads were assigned into three sequence bins as indicated. (3) Paired reads with one or both partners mapping to unannotated genomic regions were clustered based on overlapping alignments. (4) Clusters were filtered to remove singleton, stacked, and duplicate reads. (5) To determine a consensus pseudogene annotation, clusters were scanned through the Yale and ENCODE pseudogene databases as well as analyzed with a BLAT-based custom homology search. Data from individual samples were then compared to generate pseudogene expression signatures. Clusters not assigned at this stage were categorized as other potentially nonpseudogene transcripts.

See also Figures S1, S2, and S3 and Tables S1 and S2.

evidence of pseudogene transcripts obtained from disparate gene expression platforms, including public mRNA and EST databases, cap analysis gene expression (CAGE) studies, and gene identification signature-paired end tags (GIS-PET) (Ruan et al., 2007). Given the essentially anecdotal observations of pseudogene expression, only 160 expressed human pseudogenes are currently documented in ENCODE. Though this could be due to a general lack of transcription of pseudogenes, as generally presumed, it may also be reflective of an insufficient and uneven depth of coverage afforded by early gene expression analysis tools.

In this context, the recent maturation of next-generation high-throughput sequencing platforms provides unprecedented access to genome-wide expression analyses previously not achievable (Han et al., 2011a; Morozova et al., 2009). Here, we analyzed a compendium of RNA-Seq transcriptome data specifically focusing on pseudogene transcripts from a total of 293 samples encompassing 13 different tissue types, including 248 cancer and 45 benign samples. In order to carry out a systematic analysis of pseudogene expression, we developed a bioinformatics pipeline focused on detecting pseudogene transcription. This integrative approach provided evidence of expression for 2,082 distinct pseudogenes, which displayed lineage-specific, cancer-specific, as well as ubiquitous expression patterns. Taken together, this Resource nominates a multitude of expressed pseudogenes that merit further investigation to determine their roles in biology and in human disease.

RESULTS

Development of a Bioinformatics Platform for the Analysis of Pseudogene Transcription

Paired-end RNA-Seq data from a compendium of 293 samples, representing both cancer and benign samples from 13 different tissue types recently generated in our laboratory, was utilized to build a pseudogene analysis pipeline (Figure 1 and Figure S1 and Table S1 available online). Sequencing reads were mapped to the human genome (hg18) and University of California Santa Cruz (UCSC) Genes using Efficient Alignment of Nucleotide Databases (ELAND) software of the Illumina Genome Analyzer Pipeline (Table S2). Reads showing mismatches to the reference genes but mapping perfectly to unannotated regions elsewhere in the genome were used as the primary data for pseudogene expression analysis. Two or more unique, high-quality overlapping reads nucleating at the loci of differences between wild-type genes and pseudogenes were used to define de novo "clusters" (ranging from 40 to 5,000 bp). These clusters were employed for gene expression analyses in a way analogous to the "probes" used in microarray gene expression studies, though unlike predesigned and fixed probes used in microarrays, the sequence clusters used here were formed de novo, solely based on the presence (and levels) of transcripts. Thus, one or more clusters (like one or more probes in microarrays) represented a transcript, whereas the number of reads mapping to a cluster (analogous to fluorescence intensity due to probe hybridization on microarrays) provided a measure of expression of the corresponding (pseudo)genes. For example, Figure 2 shows a schematic representation of the cluster alignments for two representative pseudogenes, *ATP8A2-Ψ* (Figure 2A) and *CXADR-Ψ* (Figure 2B). As can be seen, mutation-dense regions in the reference sequence provide foci of pseudogene-specific cluster formation. Naturally, pseudogenes with sparse and dispersed mutations nucleate fewer clusters and require higher depth of coverage for reliable detection.

Overall, 2,156 unique pseudogene transcript clusters were identified, and their genomic coordinates (start and end points) were compared with the coordinates of pseudogenes annotated in the ENCODE (Zheng et al., 2007) and Yale pseudogene resources (http://www.pseudogene.org) (Karro et al., 2007), the two most comprehensive pseudogene annotation databases. Genomic coordinates of 934 unique pseudogene transcript clusters in our data set were found to overlap with the pseudogene coordinates annotated in *both* Yale and ENCODE databases. In addition, 585 clusters overlapped with Yale and 92 with ENCODE databases, displaying a high degree of overall concordance between our data and the authentic resources and highlighting a level of difference between the two reference databases (that necessitated our consideration of both resources). Further, as multiple clusters can sometimes represent one distinct pseudogene transcript, the 2,156 transcript clusters provided evidence for 2,082 distinct transcripts. Of these, 1,506 transcripts overlap with the genomic coordinates of pseudogenes in Yale and/or ENCODE, and up to 576 transcripts are potentially novel (described below) (Figure S2A). The 2,082 pseudogene transcripts, in turn, correspond to 1,437 wild-type genes, clearly indicating that the transcripts of multiple pseudogenes arisen from the same wild-type genes are also detected in our compendium. Taken together, our study provides evidence of widespread transcription of pseudogenes unraveled by high-throughput transcriptome sequencing (Table S3).

Pseudogene clusters across the sample-wise compendium reveal that pseudogenes of housekeeping genes such as ribosomal proteins are widely expressed across tissue types. Additionally, pseudogene transcripts corresponding to *CALM2* (calmodulin 2 phosphorylase kinase, delta), *TOMM40* (translocase of outer mitochondrial membrane 40), *NONO* (non-POU domain-containing, octamer-binding), *DUSP8* (dual-specificity phosphatase 8), *PERP* (TP53 apoptosis effector), and *YES* (v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1), etc. were observed in more than 50 samples each, which were further validated by pseudogene-specific RT-PCR followed by Sanger sequencing (Table S4).

Further, because our RNA-Seq compendium comprises 35- to 45-mer short sequence reads that largely generated short sequence clusters not optimal for available pseudogene analysis tools such as Pseudopipe (Zhang et al., 2006) and Pseudofam (Lam et al., 2009) used in generating ENCODE and Yale databases, we carried out a direct query of individual clusters against the human genome (hg18) using the BLAT tool from UCSC, which is ideally suited for short sequence alignment searches (Kent, 2002). Based on this "custom" analysis, or simply BLAT (Figure S2A), we were able to independently assign 1,888 clusters representing 1,820 unique pseudogenes to unique genomic locations.

Detection of Potentially Novel Pseudogene Transcripts

Comparing the genomic locations of the pseudogene clusters identified by BLAT analysis to those identified by Yale and ENCODE databases (Figure S2A), 762 clusters were found to be common to all three resources, but a remarkably large set of 585 clusters was uniquely defined by BLAT analysis alone. Some of the pseudogene transcripts thus identified included *BAT1*, *BTBD1*, *COX7A2L*, *CTNND1*, *EIF5*, *PAPOLA*, *PARP11*, *SYT*, *ZBTB12*, and others (n = 25) and were validated by Sanger sequencing (Table S4). Thus, analysis of RNA-Seq data provided a reliable assessment of expressed pseudogenes.

Though designating the BLAT-based pseudogene clusters as novel pseudogenes must await further sequence characterization (such as analysis of ORF structure and potential genesis of novel protein-coding gene family members, etc.), a small subset of clusters was seen to be localized in the vicinity of known pseudogenes. Thus, we found 92 clusters that resided adjacent (within 5 kb) to previously annotated pseudogenes (Figure S2B, left), and we hypothesize that these may represent pseudogenes with inaccurate annotations in the current databases. For example, the chromosomal coordinates of *CENTG2-Ψ* (OTTHUMT00000085288, Havana processed pseudogene) are defined in ENCODE as Chr1:177822463-177824935. As expected, we observed a cluster mapping to this locus; however, interestingly, we also observed a distinct cluster (Chr1:177825028-177826295) less than 100 base pairs away. Although unannotated in the current databases, the sequence of this adjacent locus shows a high degree of homology to the *CENTG2* parental gene (Figure S2B, right), strongly suggesting that this cluster represents an extension of the existing genomic coordinates of *CENTG2-Ψ* annotation. Similar observations were made with *HNRNPA1* and the *HNRNPA1-Ψ* on Chr6q27 (Figure S2B, right). 493 BLAT derived clusters that were not in close proximity to annotated pseudogenes likely represent putative pseudogenes currently missing in the database annotations (Table S3B).

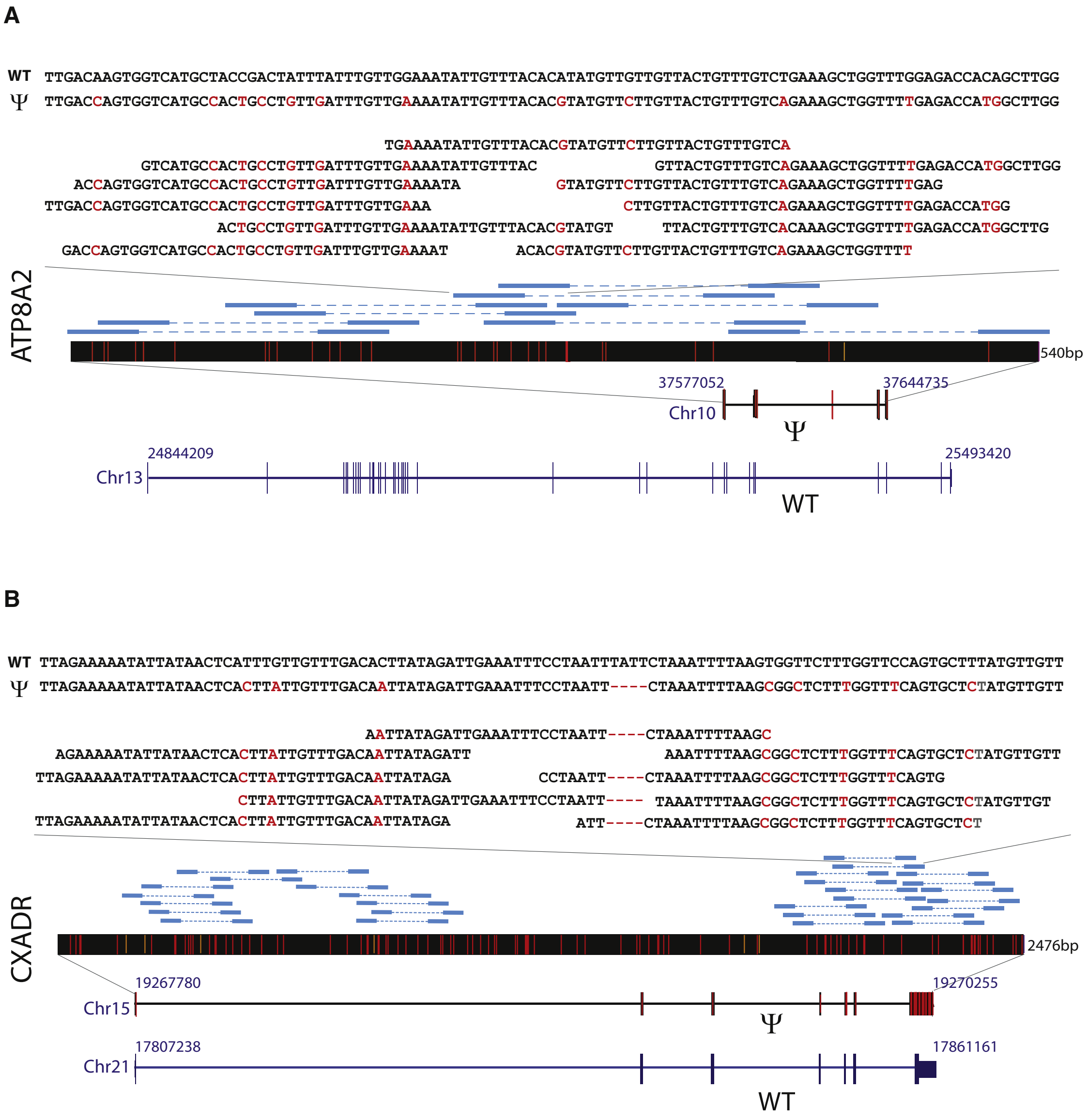

A
WT TTGACAAGTGGTCATGCTACCGACTATTTATTTGTTGGAAATATTGTTTACACATATGTTGTTGTTACTGTTTGTCTGAAAGCTGGTTTGGAGACCACAGCTTGG
Ψ TTGACCAGTGGTCATGCCACTGCCTGTTGATTTGTTGAAAATATTGTTTACACGTATGTTCTTGTTACTGTTTGTCAGAAAGCTGGTTTTGAGACCATGGCTTGG
TGAAAATATTGTTTACACGTATGTTCTTGTTACTGTTTGTCA
GTCATGCCACTGCCTGTTGATTTGTTGAAAATATTGTTTAC
GTTACTGTTTGTCAGAAAGCTGGTTTTGAGACCATGGCTTGG
ACCAGTGGTCATGCCACTGCCTGTTGATTTGTTGAAAATA
GTATGTTCTTGTTACTGTTTGTCAGAAAGCTGGTTTTGAG
TTGACCAGTGGTCATGCCACTGCCTGTTGATTTGTTGAAA
CTTGTTACTGTTTGTCAGAAAGCTGGTTTTGAGACCATGG
ACTGCCTGTTGATTTGTTGAAAATATTGTTTACACGTATGT
TTACTGTTTGTCACAAAGCTGGTTTTGAGACCATGGCTTG
GACCAGTGGTCATGCCACTGCCTGTTGATTTGTTGAAAAT
ACACGTATGTTCTTGTTACTGTTTGTCAGAAAGCTGGTTTT
ATP8A2
540bp
37577052
37644735
Chr10
Ψ
24844209
25493420
Chr13
WT
B
WT TTAGAAAAATATTATAACTCATTTGTTGTTTGACACTTATAGATTGAAATTTCCTAATTTATTCTAAATTTTAAGTGGTTCTTTGGTTCCAGTGCTTTATGTTGTT
Ψ TTAGAAAAATATTATAACTCACTTATTGTTTGACAATTATAGATTGAAATTTCCTAATT----CTAAATTTTAAGCGGCTCTTTGGTTTCAGTGCTCTATGTTGTT
AATTATAGATTGAAATTTCCTAATT----CTAAATTTTAAGC
AGAAAAATATTATAACTCACTTATTGTTTGACAATTATAGATT
AAATTTTAAGCGGCTCTTTGGTTTCAGTGCTCTATGTTGTT
TTAGAAAAATATTATAACTCACTTATTGTTTGACAATTATAGA
CCTAATT----CTAAATTTTAAGCGGCTCTTTGGTTTCAGTG
CTTATTGTTTGACAATTATAGATTGAAATTTCCTAATT----
TAAATTTTAAGCGGCTCTTTGGTTTCAGTGCTCTATGTTGT
TTAGAAAAATATTATAACTCACTTATTGTTTGACAATTATAGA
ATT----CTAAATTTTAAGCGGCTCTTTGGTTTCAGTGCTCT
CXADR
2476bp
19267780
19270255
Chr15
Ψ
17807238
17861161
Chr21
WT

Figure 2. Schematic Representation of Cluster Alignments with Pseudogene Transcripts

(A and B) The relative genomic structures of the parental genes are shown aligned to the respective pseudogenes, with their chromosomal locations indicated on the sides, (A) *ATP8A2*-Ψ and (B) *CXADR*-Ψ. The sequencing alterations distinguishing the pseudogene from the parental gene are indicated in red. The pseudogene transcripts are illustrated as black bars with red hatches, which indicate divergence from the parental sequence, and the length of the transcript in base pairs is shown on the side. These representations are then overlaid with schematics of paired-end reads used to form pseudogene clusters (in blue), followed by overlapping sequences in a zoomed-in region of the cluster. A comparative representation of the parental (WT) and pseudogene (Ψ) sequences for the specified region is shown on top.

See also Figure S4.

Next, we assessed the technical and analytical factors influencing the yield of pseudogene transcripts. As may be expected, a positive correlation was observed between the sequencing depth and total number of pseudogene transcripts (correlation coefficient, +0.65) (Figure S3A). However, no significant correlation was observed between the absolute measure of percent similarity between pseudogene-WT pairs and pseudogene yield. Importantly, the metric of overall percent similarity accounts for gap penalty and mismatches in BLAT search, but it is the "distribution" of the mismatches that is critical in resolving pseudogenes from nearly identical wild-type sequences; for example, a few mismatches, accumulated in a small stretch, are more effective in confidently distinguishing pseudogene expression from wild-types as compared to a higher number of mismatches that are scattered over long stretches of sequence (Figure 2). Thus, three primary factors determine the detection of pseudogene transcription by RNA-Seq: (1) the level of expression of the pseudogenes (i.e., the higher the level of expression, the higher the likelihood of detection), (2) the depth of RNA sequencing, and (3) overall distribution of mismatches with respect to the wild-type.

To explore the loci of transcription regulatory elements associated with pseudogene transcription, we carried out ChIP-Seq analysis of a breast cancer cell line MCF7 probed with H3K4me3, a histone mark associated with transcriptionally active chromosomal loci, and integrated the results with the MCF7 pseudogene transcript data. Interestingly, we observed a statistically significant enrichment of H3K4me3 peaks at expressed pseudogene loci as compared to nonexpressed pseudogenes (p = 0.0054) (Figure S3B), suggesting that the pseudogene transcripts observed by RNA-Seq are associated with transcriptionally active genomic loci. Interestingly, the pseudogene transcripts associated with H3K4me3 peaks encompass both unprocessed and processed pseudogenes, with no discernible differences in the pattern of expression. Considering the role of 3′ UTRs with MREs in ceRNA regulatory networks, we also looked at the frequency of 3′ UTR sequences retained in our set of pseudogene transcripts and observed that at least 71% of all pseudogene transcripts retain distinct 3′ UTR sequences similar to their cognate wild-type genes (Figure S3C). Interestingly, comparing the pseudogene transcripts with a list of genes implicated in ceRNA networks (Han et al., 2011b; Tay et al., 2011), we observed more than 400 overlapping transcripts (Table S5). The presence of noncoding pseudogene transcripts with similar 3′ UTRs (and MREs) adds a further level of complexity to ceRNA regulatory networks.

Next, we assessed a potential correlation between the expression of pseudogenes present within the introns of unrelated, expressed genes with their "host" genes. Interestingly, no significant association was observed, suggesting that pseudogenes are likely subject to independent regulatory mechanisms even when residing within other transcriptionally active genes. Further, our observations with the breast-specific unprocessed pseudogene *ATP8A2* (likely arisen from duplication of wild-type *ATP8A2*, thus likely harboring similar promoter elements) also indicate that there is no apparent correlation between the pseudogene expression with the wild-type gene that is expressed ubiquitously (described later). Thus, in summary, although it is tempting to speculate that pseudogene expression may be regulated by the promoter elements from the cognate gene or the host genes, our data suggest that more complex/indirect factors may be at play. Next, we assessed a possible correlation between the expression of pseudogenes with that of cognate wild-type genes, and intriguingly, no significant pattern of correlation was observed (Figure S3D).

Focusing on the pseudogenes whose genomic coordinates are annotated in the reference databases, we next analyzed the expression profiles of the 1,056 unique transcripts.

Patterns of Pseudogene Expression in Human Tissues

Analyzing the expression data from 248 cancer and 45 benign samples from 13 different tissue types (total 293 samples), we observed broad patterns of pseudogene expression, including 1,056 pseudogenes that were detected in multiple samples (Table S6), which supports the hypothesis that transcribed pseudogenes contribute to the typical transcriptional repertoire of cells. In addition, we identified distinct patterns of pseudogene expression, akin to that of protein-coding genes, including 154 highly tissue/lineage-specific and 848 moderately tissue/lineage-specific (or enriched) pseudogenes (Figure 3A). Moreover, we found 165 pseudogenes exhibiting expression in more than 10 of the 13 tissue types examined, and these we classified as ubiquitous pseudogenes whose transcription is characteristic of most cell types (Figure 3A, bottom).

Of the 165 ubiquitous pseudogenes, a majority belonged to housekeeping genes, such as glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*), ribosomal proteins, several cytokeratins, and other genes widely expressed in most cell types. This is expected, as these genes are known to have numerous pseudogenes, and it is likely that several of these pseudogenes retain the capacity for widespread transcription, mimicking their protein-coding counterparts.

A second set of pseudogenes exhibited near ubiquitous expression but were frequently transcribed at lower levels in most tissues and robustly transcribed in one or two tissues. These pseudogenes were termed "nonspecific," and this group harbors more than 870 pseudogenes, comprising a large portion of our data set (Figure 3A, middle). Many of the pseudogenes previously shown to be expressed were found in this category, including some pseudogenes reported as tissue specific, such as *CYP4Z2P*, a pseudogene previously reported to be expressed only in breast cancer tissues (Rieger et al., 2004). Other candidates observed in this category include pseudogenes derived from *Oct-4* (Kastler et al., 2010), *Connexin-43* (Bier et al., 2009; Kandouz et al., 2004), and *BRAF* (Zou et al., 2009), among others (Table S6).

Though powerful, our approach is nevertheless limited to pseudogene transcripts that are expressed above the current threshold of detection by RNA-Seq and possess distinct stretches of sequence mismatches compared with their protein-coding parental genes. Thus, for example, *PTENP1*, a pseudogene of *PTEN* recently implicated in the biology of the phosphatidylinositol 3-kinase (PI3K) signaling pathway, was not detected in our compendium possibly due to the preponderance of cancer samples in our cohort, which tend to show low expression or deletion of this pseudogene (Poliseno et al., 2010).

Lineage- and Cancer-Specific Pseudogene Expression Signatures

Lineage-specific pseudogene transcripts may have the potential for lineage-specific functions and may represent novel elements that facilitate biological characteristics that are unique to distinct tissue types. In this regard, we observed 154 pseudogenes with highly specific expression patterns, including pseudogenes derived from *AURKA* (kidney samples), *RHOB* (colon samples), and *HMGB1* (myeloproliferative neoplasms [MPNs]) (Figure 3A, top). Interestingly, however, lineage-specific pseudogenes tended to represent a small fraction of all pseudogenes expressed in a given tissue type, and the total number of lineage-specific pseudogenes observed in a tissue type did not show a correlation with the total number of samples analyzed. For example, B-lymphocyte cells (n = 19) and MPNs (n = 9) showed more lineage-specific pseudogenes than breast (n = 64) or prostate (n = 89). Conversely, we did observe more pseudogene

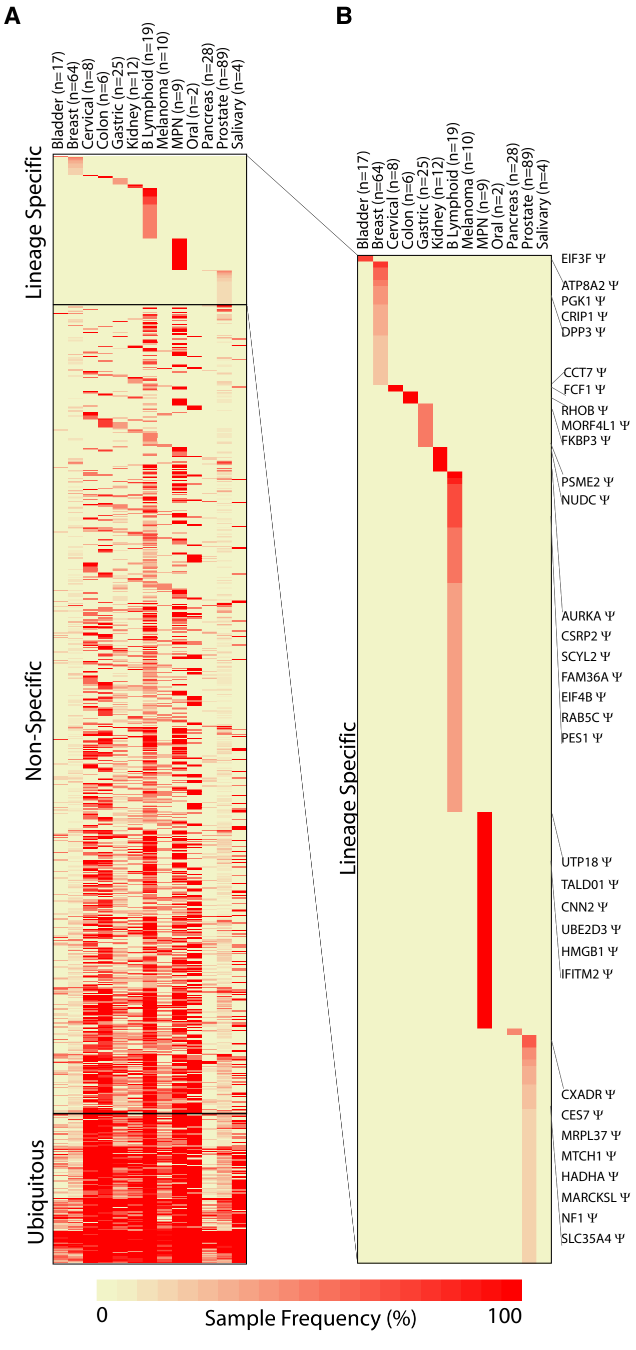


Figure 3. Tissue/Lineage-Specific Pseudogene Expression Profiles

(A) Heatmap of pseudogene expression sorted on the basis of tissue-specific expression displays tissue-specific (top), tissue-enriched/nonspecific (middle), and ubiquitously expressed pseudogenes (bottom).

(B) Zoomed-in version of the top panel displaying tissue-specific expressed pseudogenes. The columns represent different tissues, with the number of samples in parentheses. The rows represent individual clusters mapping to specific pseudogenes. The color intensity represents the frequency (%) of samples in a tissue type showing expression of a given pseudogenes (according to the scale indicated at the bottom). The key clusters are labeled with their corresponding parental gene symbols. MPN, myeloproliferative neoplasms.

See also Table S6.

transcripts in samples with longer read lengths and deeper coverage, as expected. Together, these data both confirm and formalize previous anecdotal observations of lineage-specific pseudogene expression patterns by exploiting the power of RNA-Seq to resolve individual transcripts (Figure 3B) (Bier et al., 2009; Lu et al., 2006; Rieger et al., 2004; Zou et al., 2009).

Because our sample compendium has a substantial number of cancer samples, we next focused on pseudogenes with cancer-specific expression. Though a majority of the pseudogenes examined were found in both cancer and benign samples, we observed 218 pseudogenes expressed only in cancer samples, of which 178 were observed in multiple cancers and 40 were found to have highly specific expression in a single cancer type only (Figure 4A and Table S7). Consistent with our previous results (Figure 3), we found that the number of cancer-type-specific pseudogenes did not correlate with the number of samples sequenced in a given cancer type. These results suggest that cancer samples harbor transcriptional patterns of pseudogenes that are both lineage and cancer specific.

Among the cancer-specific pseudogenes, a few noteworthy examples included pseudogenes derived from the eukaryotic translation initiation factors *EIF4A1* and *EIF4H*, the heterogeneous nuclear ribonucleoprotein *HNRPH2*, and the small nuclear ribonucleoprotein *SNRPG* (Figure 4B). Moreover, we observed pseudogenes corresponding to known cancer-associated genes, including *RAB-1*, a Ras-related protein; *VDAC1*, a type-1 voltage-dependent anion-selective channel/porin; *RCC2*, a regulator of chromosome condensation 2; and *PTMA*, prothymosin alpha. Interestingly, the parental protein-coding *PTMA* gene has given rise to five processed pseudogenes that retain consensus TATA elements, individual transcriptional start sites, and intact open reading frames that may potentially code for proteins closely related to the parental PTMA protein. Importantly, we find expression of *PTMA*-derived pseudogenes in more than 30 cancer samples, but not in any benign cells, and these data suggest that *PTMA*-derived pseudogenes may not only contribute transcripts to cancer cell biology but potentially proteins as well, warranting further study of these pseudogenes in tumorigenesis.

Prostate Cancer Pseudogenes

To investigate individual pseudogenes in greater detail, we focused on pseudogenes associated with prostate and breast cancer, as our compendium has a substantial number of these two cancer types represented. Analysis of lineage-specific pseudogenes restricted to prostate cancers identified numerous pseudogenes, including several derived from parental genes known to be altered or dysregulated in cancer; for example, *NDUFA9*, which encodes an *NADH* oxidoreductase component of mitochondrial complex I that is reported to be upregulated in testicular germ cell tumors (Dormeyer et al., 2008); *EPCAM*, an epithelial cell adhesion molecule involved in cancer and stem cells signaling (Munz et al., 2009); and *CES7*, known to be expressed only in the male reproductive tract (Gang et al., 2011) (Figure 3B and Table S6). Among the prostate cancer specific pseudogenes, *CXADR*-Ψ, a processed pseudogene on chromosome 15, was of immediate interest, as the parental *CXADR* protein demonstrates putative tumor suppressor functions and

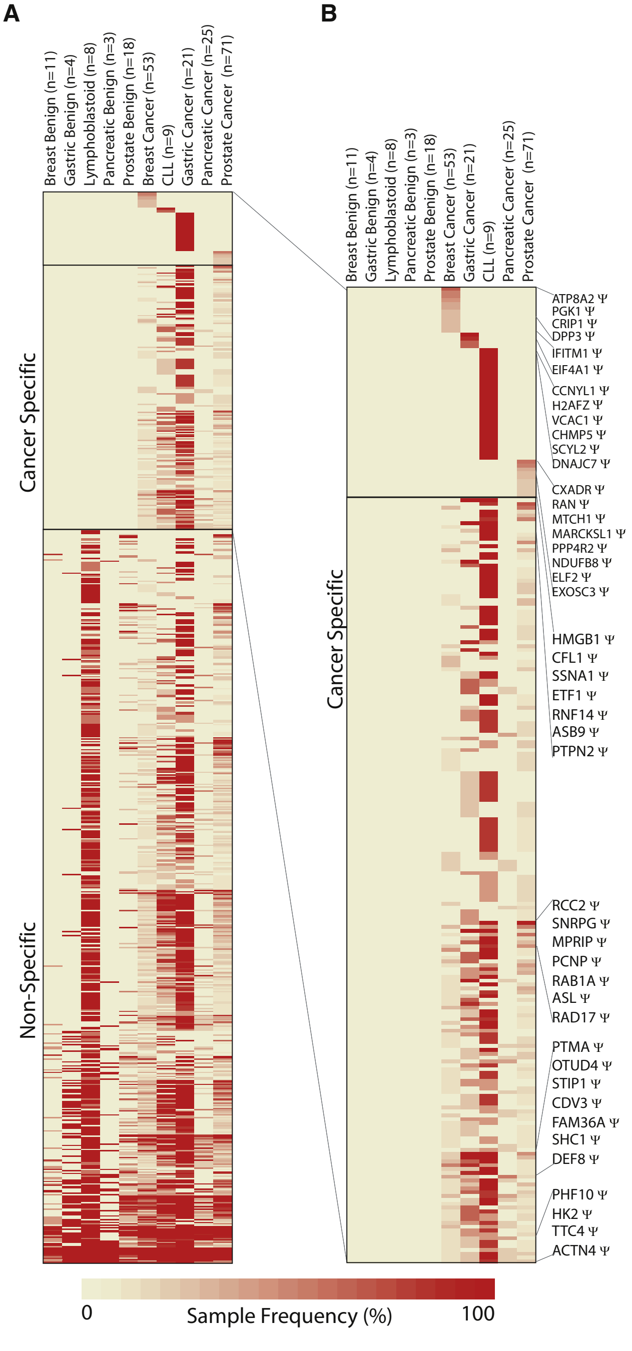


Figure 4. Cancer-Specific Pseudogene Expression Profiles

(A) Heatmap of pseudogene expression sorted according to cancer-specific expression patterns displays pseudogene transcripts specific to individual cancers (top), common across multiple cancers (tissue-enriched; middle), and nonspecific (bottom).

(B) Zoomed-in version of the top panel displaying individual cancer-specific expressed pseudogenes. The columns represent different tissues with the number of samples in parentheses. The rows represent individual clusters mapping to specific pseudogenes. The color intensity represents the frequency (%) of samples in a tissue type showing expression of a given pseudogenes (according to the scale indicated at the bottom). The key clusters are labeled with their corresponding parental gene symbols.

See also Figure S6 and Table S7.

its loss is implicated in α-catenin silencing (Pong et al., 2003). We therefore selected this pseudogene for further study in prostate cancer and first evaluated custom Taqman assays that could distinguish *CXADR*-Ψ from parental *CXADR*. The expression levels showed strong correlation with the RNA-Seq data (Figure S3E). *CXADR*-Ψ expression was found to be upregulated in ~25% of prostate cancer tissues, with minimal expression seen in benign prostate samples and nonprostate tissues (Figure 5A). No correlation was observed between *CXADR*-Ψ and parental *CXADR* expression, although parental *CXADR* also had some proclivity for prostate cancer-specific expression (Figure 5B). Interestingly, *CXADR*-Ψ expression was nearly restricted to prostate cancers lacking an ETS gene fusion, with few ETS-positive samples exhibiting expression of this pseudogene. By contrast, parental *CXADR* gene expression was found in both ETS-positive and ETS-negative samples (Figure 5C). Finally, we interrogated *CXADR*-Ψ and *CXADR* parental gene expression in a set of six prostate patients with matched cancer and benign tissues (including four ETS-negative and two ETS-positive pairs). Again, ETS-negative prostate cancer samples displayed marked upregulation of *CXADR*-Ψ compared to the ETS-positive patients, with parental *CXADR* expression being fairly constant between this set of patients (Figure 5D). To establish the expression of *CXADR*-Ψ transcript, we were able to clone *CXADR*-Ψ cDNA from two RNA-Seq-positive prostate cancer samples (Figure S5A), and as predicted, these clones showed perfect sequence similarity to the pseudogene *CXADR*-Ψ and only 84% to *CXADR* wild-type gene (Figure S5B).

In the course of these analyses, we also identified a prostate-cancer-specific readthrough transcript involving *KLK4*, an androgen-induced gene, and *KLKP1*, an adjacent pseudogene. This chimeric RNA transcript *KLK4-KLKP1*, combining the first two exons of *KLK4* with the last two exons of *KLKP1*, retains an open reading frame incorporating 54 amino acids encoded by the *KLKP1* pseudogene in the putative chimeric protein (Figure S6A). Curiously, this readthrough was recently described in the prostate cancer cell line LNCaP as a *cis* sense-antisense chimeric transcript (Lai et al., 2010). Intriguingly, the *KLK4-KLKP1* transcript was highly expressed in 30%–50% of prostate cancer tissues, and this expression was lineage and cancer specific, with minimal expression seen in benign prostate and other tissues (Figure S6B). These data suggest that the *KLK4-KLKP1* may warrant further study as a potential biomarker of prostate cancer as well as a candidate protein implicated in the biological complexity of this disease.

Breast Cancer Pseudogenes

Among the pseudogene candidates in breast cancer, we identified a unprocessed pseudogene cognate to *ATP8A2*, a LIM domain-containing protein speculated to be associated with stress response and proliferative activity (Khoo et al., 1997) (Figure 3A, top, and Table S3). Because *ATP8A2*-Ψ on chromosome 10 displays substantial sequence divergence from the cognate *ATP8A2*-WT gene on chromosome 13, it lends high confidence to our computational identification, and we selected this candidate for further validation. Taqman assays distinguishing *ATP8A2*-WT transcripts from *ATP8A2*-Ψ showed a strong correlation (r^2 = 0.98) with the expression pattern obtained

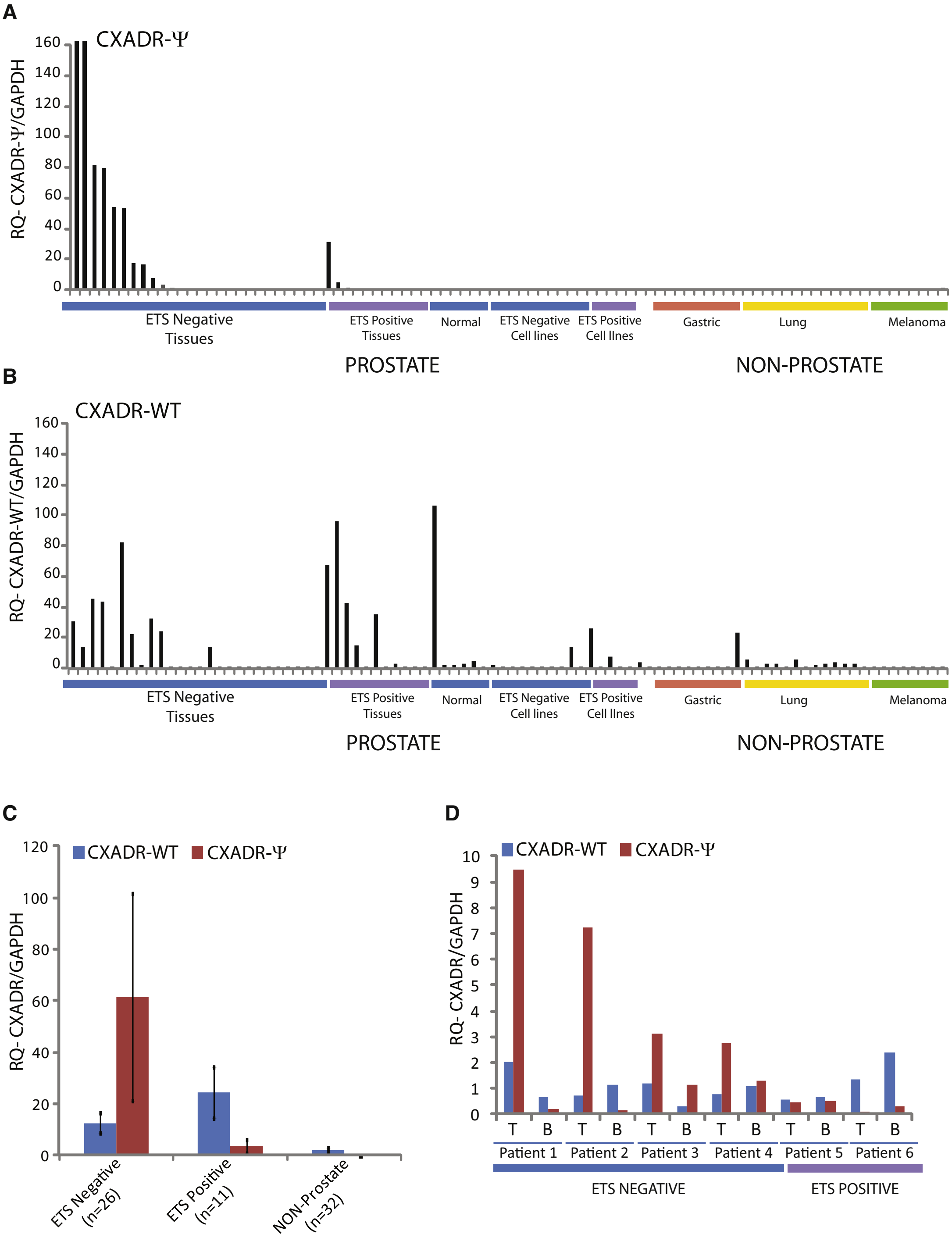


Figure 5. Expression of *CXADR*-Ψ in Prostate Cancer

(A and B) Histogram of expression values (y axis) of *CXADR*-Ψ (A) and *CXADR*-WT (B) across a panel of tissue samples (x axis). The order of samples on the x axis is identical in both graphs to facilitate a visual comparison.

(C) A summary histogram of the expression values of *CXADR*-Ψ and *CXADR*-WT in prostate cancers either harboring or lacking an ETS transcription factor gene fusion or in nonprostate samples.

(D) Expression of *CXADR*-Ψ and *CXADR*-WT in matched pairs of tumor and benign samples from prostate cancer patients. The patients' ETS status is indicated by the bar below.

T, prostate cancer; B, matched benign adjacent prostate. The expression values were normalized against *GAPDH*. Error bars represent means ± SE of the mean. See also Figure S5.

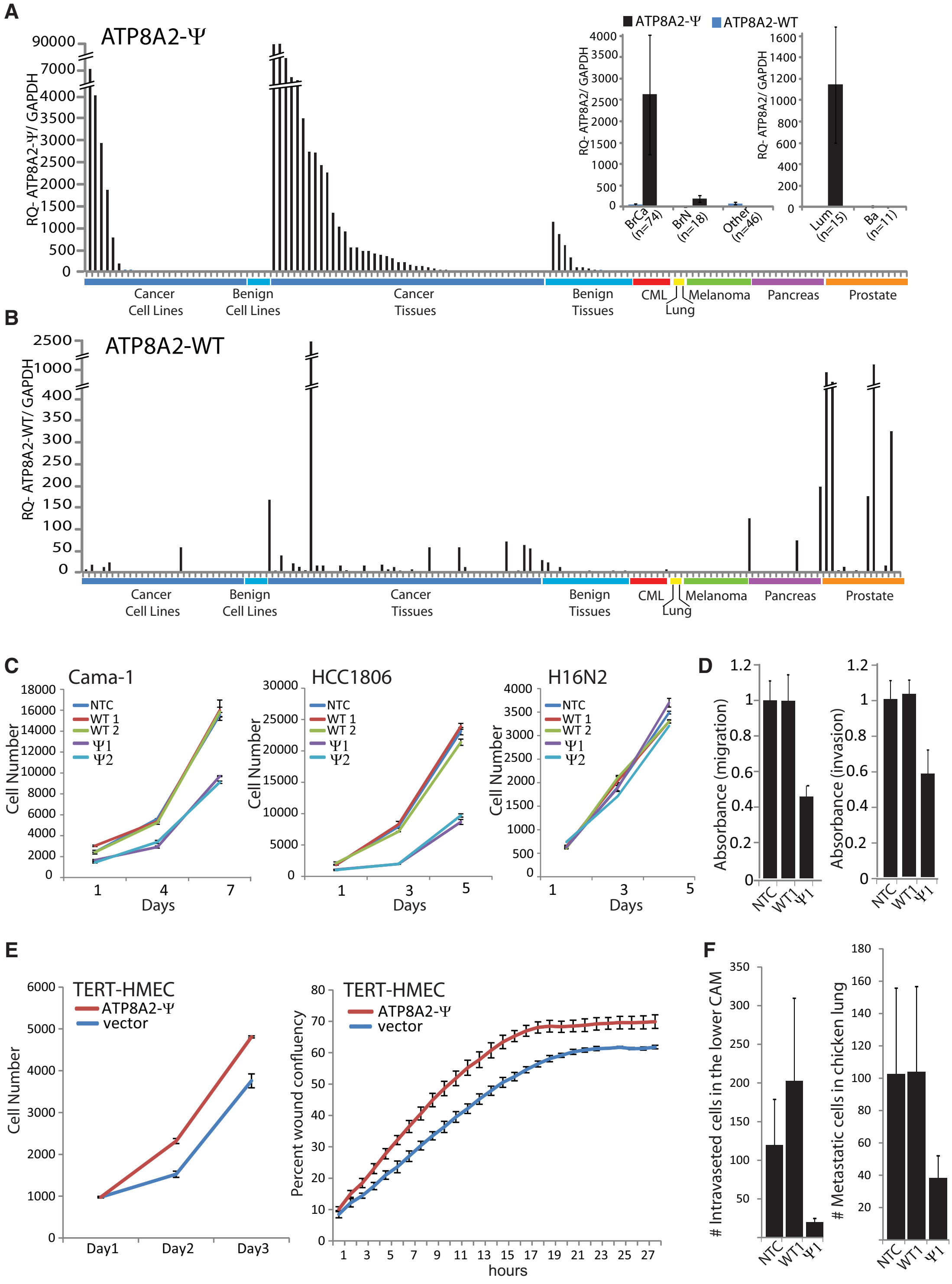


Figure 6. Expression of *ATP8A2-Ψ* in Breast Cancer

(A and B) Histogram of expression values (y axis) of *ATP8A2-Ψ* (A) and *ATP8A2*-WT (B) across a panel of tissue samples (x axis). The order of samples on the x axis is identical in both graphs to facilitate a visual comparison. (Inset) A summary histogram of the expression values of *ATP8A2-Ψ* and *ATP8A2*-WT in breast cancer samples relative to benign breast and other tissues (left) and luminal versus basal breast cancer subtypes (right). The expression values were normalized against *GAPDH*.

(C) Cell proliferation assays following siRNA knockdowns of *ATP8A2*-WT and -Ψ as indicated. NTC, nontargeting control; WT, siRNA against wild-type *ATP8A2*; Ψ, siRNA against *ATP8A2-Ψ*.

by RNA-Seq (Figure S3E), with *ATP8A2*-Ψ expression found to be restricted to breast samples, the highest levels seen in a subset of breast cancer tissues and cell lines (Figures 6A and 6B). By contrast, *ATP8A2*-WT expression was highly variable across different tissue types and showed no correlation with *ATP8A2*-Ψ expression (Figure 6B).

We were further intrigued by the pattern of *ATP8A2*-Ψ expression within breast tumors, where ~25% of tumors demonstrate extremely high levels of this pseudogene, suggesting that *ATP8A2*-Ψ may contribute to a particular subtype of breast cancer. We therefore analyzed *ATP8A2*-Ψ expression with respect to luminal and basal breast subtypes, two prominent categories of breast cancer with distinct molecular and clinical characteristics. Unexpectedly, we found that *ATP8A2*-Ψ expression was restricted to tumors with luminal histology, whereas basal tumors showed minimal expression of this pseudogene (Figure 6A, right). The wild-type *ATP8A2* transcript did not display this pattern of expression.

To investigate a potential role of *ATP8A2*-Ψ expression in breast cancer, first we carried out siRNA-based knockdown of both the wild-type and pseudogene RNA in two independent breast cancer cell lines that expressed both the transcripts (Figure S7A). Knockdown of *ATP8A2*-Ψ with two independent siRNAs was found to specifically inhibit the proliferation of overexpressing cell lines Cama-1 and HCC1806 (Figure 6C), but not the cell lines with no detectable levels of *ATP8A2*-Ψ, for example, the benign breast epithelial cell line H16N2 (Figure 6C, right) and a pancreatic cancer cell line, BXPC3 (Figure S7D). Knockdown of *ATP8A2*-Ψ (but not *ATP8A2*-WT) also resulted in reduced cell migration and invasion seen in in vitro Boyden Chamber assays (Figure 6D) as well as in in vivo intravasation and metastasis in chicken chorioallantoic membrane xenograft assay (Figure 6F). In contrast, knockdown of wild-type *ATP8A2* had no effect on the proliferation of any of the cell lines tested, suggesting an unexpected growth regulatory role for *ATP8A2*-Ψ (Figure 6C). Surprisingly, though the knockdown of wild-type *ATP8A2* had a minimal effect on the pseudogene transcript levels, *ATP8A2*-Ψ-specific siRNAs, apart from reducing the *ATP8A2*-Ψ transcript, also reduced the wild-type protein levels (Figures S7C and S7E). Thus clearly, unlike *Oct4* and *BRAF* pseudogene transcripts having an inverse correlation with the wild-type transcript levels, *ATP8A2*-Ψ and wild-type *ATP8A2* transcripts (Figures 6A and 6B) and protein (Figure S7E) do not seem to be regulated in this manner. Subsequently, to assess the phenotypic effect of *ATP8A2*-Ψ overexpression in benign cells, we cloned and overexpressed the full-length *ATP8A2* pseudogene cDNA in benign breast epithelial cell line TERT-HMEC. Two independent pooled populations of *ATP8A2*-Ψ-overexpressing TERT-HMEC cells were found to undergo increased proliferation and migration (Figure 6E), indicating the potential oncogenic nature of this breast-specific pseudogene transcript.

DISCUSSION

The recent advances in high-throughput transcriptome sequencing have revealed widespread expression of noncoding RNAs in the context of development and differentiation (Khachane and Harrison, 2010; Nagalakshmi et al., 2008; Pickrell et al., 2010; Prensner et al., 2011; Wilhelm et al., 2008). These studies, however, do not include pseudogene expression analyses in their purview, likely due to the challenge of extremely close sequence similarity with wild-type cognate genes. Here, we interrogated the potential of RNA-Seq data to unambiguously detect pseudogene transcripts and to assess whether pseudogene expression is more common in the transcriptome than previously realized. Surprisingly, we found evidence of a widespread expression of pseudogenes in our cancer transcriptome resource, including 1,500 pseudogenes annotated in the Yale and ENCODE databases, redefined the genomic coordinates of ~100 pseudogenes in existing databases, and nominated more than 400 potentially novel pseudogenes. In aggregate, our analysis considerably expands the spectrum of expressed pseudogenes documented previously (Harrison et al., 2005; Yao et al., 2006; Zheng et al., 2007).

The extreme sequence similarity between pseudogenes and cognate wild-type genes suggests a functional role for pseudogene transcripts; indeed, pseudogene expression has been associated with both downregulation of cognate wild-type gene, such as *eNOS* in ovary, as well as a positive effect on the expression of the wild-type gene, as demonstrated recently, wherein *PTENP1* expression upregulates *PTEN* expression in prostate cells (Poliseno et al., 2010). Interestingly, a class of pseudogenes called ‘‘unitary pseudogenes’’ does not have extant cognate wild-type genes (Zhang et al., 2010). Nevertheless, as most pseudogenes do have distinct cognate wild-type genes, we assessed the correlation between expressed pseudogenes and their cognate wild-type genes across multiple samples (of the same tissue type or across diverse tissue types) and did not observe a statistically significant correlation. This is not surprising, partly because our data set is comprised of a heterogeneous set of samples representing diverse tissue types. Further, the sensitivity of detection of individual pseudogene transcripts is limited by the degree and distribution of dissimilarity with the wild-type gene that determines the ‘‘effective’’ depth of coverage; this limits the number of samples showing measurable expression of individual pseudogene-wild-type pairs, making it difficult to conduct robust statistical analyses. Future studies involving larger sample sets with higher depth of coverage and longer read length may be better able to resolve this question.

Taken together, our study provides a systematic approach to analyze expressed pseudogenes using RNA-Seq data, enabling comparisons of cancer versus benign tissues in multiple solid

(D) Boyden chamber assay showing cell migration (left) and invasion through matrigel (right).

(E and F) (E) The effect of *ATP8A2*-Ψ overexpression in TERT-HMEC cells on cell proliferation (left) and cell migration based on Incucyte wound confluency assay (right) and (F) chicken chorioallantoic membrane assay of HCC-1806 cells treated with nontargeting control siRNA, *ATP8A2*-WT, or *ATPA2*-Ψ siRNA showing relative number of cells intravasated in the lower CAM (left) and metastatic cells in chicken lung (right).

Error bars represent means ± SE of the mean.

tumors. Our efforts lend additional credence to the capacity of RNA-Seq to "re-define" the functional elements of the genome and "re-annotate" the population of pseudogenes implicated in human cell biology. Our approach overcomes the limitations of previous analyses of pseudogene expression, which were primarily anecdotal and heterogeneous in nature, and our methodologies suggest avenues to reconcile the difficulty in distinguishing pseudogene expression from parental protein-coding gene expression—a facet that is important for all RNA-Seq studies aiming to provide an accurate picture of gene expression. Finally, we describe *ATP8A2*-Ψ and *CXADR*-Ψ pseudogenes preferentially associated with distinct subsets of breast cancer and prostate cancer patients, respectively.

The recent description of intricate regulatory networks of protein-coding transcripts called competitive endogenous RNAs (ceRNAs) defined on the basis of coordinated regulation by common sets of microRNA response elements (MREs)—first intimated by Salmena et al. (Salmena et al., 2011) and subsequently supported by experimental results from multiple groups(Cesana et al., 2011; Han et al., 2011b; Karreth et al., 2011; Tay et al., 2011)—implicates potential noncoding functions for many protein-coding transcripts. In this context, pseudogene transcripts could provide an additional layer of complexity in conjunction with their cognate wild-type genes or independently.

The cancer/tissue-specific pseudogene expression signatures described here highlight the need to factor in pseudogene expression in all high-throughput gene expression studies and also show that pseudogene expression merits further exploration in its own right as an additional layer of transcriptional complexity. To facilitate further analyses, we provide here an extensive resource of RNA-Seq data of human cancer-related tissues and cell lines.

EXPERIMENTAL PROCEDURES

Data Set

Paired-end transcriptome sequence reads (2 × 40 and 2 × 80 base pairs) were obtained from a total of more than 293 samples from 13 tissue types (Figure S1 and Table S1). Each sample was sequenced on an Illumina Genome Analyzer I or II according to protocols provided by Illumina as described earlier (Palanisamy et al., 2010).

Pseudogene Analysis Pipeline

Paired-end transcriptome reads were mapped to the human genome (NCBI36/hg18) and University of California Santa Cruz (UCSC) Genes using Efficient Alignment of Nucleotide Databases (ELAND) software of the Illumina Genome Analyzer Pipeline, using 32 bp seed length and allowing up to two mismatches; detailed mapping status is represented in Table S2. Passed purity filter reads obtained from Illumina export and extended output files (as described before) were parsed and binned into three major categories: (1) both of the paired reads map to annotated genes; (2) one or both of the paired reads map to unannotated regions in the genome; and (3) neither of the reads map (these include viral, bacterial, and other contaminant reads, as well as sequencing errors). The paired reads with one or both partners mapping to an unannotated region were clustered based on overlaps of aligned sequences using the chromosomal coordinates of the clusters. Singleton reads that did not cluster or stacked\duplicated reads with the same start and stop genomic coordinates (potential PCR artifacts) were filtered out. Passed filter clusters were defined as units of transcript expression (analogous to a "probe" on microarray platforms). These clusters were screened against two human pseudogene resources, Yale human pseudogene (Build 53, http://pseudogene.org/) (Karro et al., 2007) and Gencode (October 2009, http://genome.ucsc.edu/ENCODE/) (Zheng et al., 2007), to identify and annotate pseudogene clusters. The processed, duplicated, and fragmented categories of pseudogene entries from Yale and the entries corresponding to Level 1+2 (Manual Gene Annotations) and Level 3 (Automated Gene Annotations) from Gencode were used. The clusters were also subjected to homology search using the alignment tool BLAT (http://www.soe.ucsc.edu/~kent) (Kent, 2002) for an independent annotation. Sequence reads from individual samples were queried against the resultant clusters defined by the union of Yale, ENCODE, and BLAT output to assess the expression of pseudogenes (Figure 1 and Table S3). The cutoff value for pseudogene expression in a sample was set at five or more reads mapping to at least one cluster in a putative pseudogene transcript. Pseudogene transcripts (one or more probes overlapping with either Yale or ENCODE) detected in two or more samples in a tissue type and absent in all other tissue types were defined as tissue/lineage specific. Pseudogene probes detected in 10 out of 13 samples were designated as ubiquitous. All other cases were described as an intermediate category. Pseudogene transcripts detected in three or more cancer samples and absent in all benign samples were designated as cancer specific.

We carried out multiple correlation analyses (Figure S3), including: (1) passed filter reads (sequence yield) with total number of pseudogene transcripts observed per sequencing run (pseudogene transcript coverage); (2) expression of genes and pseudogenes carried out using 173 gene-pseudogene pairs in 64 samples that each show nonzero expression in at least ten samples across the data set; (3) expression levels of *ATP8A2* and *CXADR* pseudogenes obtained from RNA-Seq and qPCR; (4) ChIP-Seq analysis of a breast cancer cell line MCF7 that was probed with H3K4me3 and compared with MCF7 pseudogene transcript data; and (5) pseudogene transcripts with 3′ UTR sequences (± 2 kb) that were compared with 3′ UTR sequences of their cognate genes using BLAT.

Pseudogene transcripts showing an overlap with transcripts involved in ceRNA network genes reported previously were tabulated (Sumazin et al., 2011 and Tay et al., 2011) (Table S5). The entire sequence data set will be submitted to dbGAP after securing requisite approvals.

RNA Isolation and cDNA Synthesis

Total RNA was isolated using Trizol and an RNeasy Kit (Invitrogen) with DNase I digestion according to the manufacturer's instructions. RNA integrity was verified on an Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA). cDNA was synthesized from total RNA using Superscript III (Invitrogen) and random primers (Invitrogen).

Quantitative Real-Time PCR

Quantitative real-time PCR (qPCR) was performed using Taqman or SYBR green-based assays (Applied Biosystems, Foster City, CA) on an Applied Biosystems 7900HT Real-Time PCR System, according to standard protocols. The Taqman assays for *CXADR* and *ATP8A2* assays were custom designed based on regions of differences between the wild-type and pseudogene sequences (Figure S4). Oligonucleotide primers for SYBR green assays were obtained from Integrated DNA Technologies (Coralville, IA). The housekeeping gene *GAPDH* was used as a loading control. Fold changes were calculated relative to *GAPDH* and normalized to the median value of the benign samples.

CXADR-Ψ_F CGGTTTCAGTGCTCTATGTTGTTTG; *CXADR*-Ψ_R TAAATTTAGGATTACATGTTTCTAGAACA; *CXADR*-Ψ_M 6FAM ATGCCATCCAAAACCA; *ATP8A2*-Ψ_F CTGGTGTTCTTTGGCATCTACTCA; *ATP8A2*-Ψ_R CAGCTCAGGATCACAGTTGCT; *ATP8A2*-Ψ_M 6FAM CTGGTCCACCATTCTC; *ATP8A2*-WT_F ATCCTATTGAAGGAGGACTCTTTGGA; *ATP8A2*-WT_R CCAGCAAATTCCCAAGGTCAGT; *ATP8A2*-WT_M 6FAM AAGGGCAGCCATTACT; *KLK4-KLKP1*_F ATGGAAAACGAATTGTTCTG; and *KLK4-KLKP1*_R CAGTGTTCCGGGTGATGCAG.

Additionally, inventoried Taqman assays for *CXADR*-WT (Hs00154661_m1) and *ATP8A2*-WT (assay ID hs00185259_m1) were used.

RT-PCR and Sanger Sequencing

Sequence stretches unique to pseudogene transcripts were identified by aligning the candidate pseudogene sequences with their corresponding wild-type genes. PCR primers specific to pseudogene transcripts (Table S4) were used to amplify pseudogene cDNAs from index samples followed by Sanger sequencing of the PCR products. The resultant sequences were analyzed using ClustalW to compare the identity between pseudogene and cognate wild-type sequences.

Cell Proliferation Assays

Experimental cells were transfected with siRNAs using oligofectamine reagent (Life Sciences), and 3 days posttransfection, the cells were plated for proliferation assays. At the indicated times, cell numbers were measured using Coulter Counter.

Wound Healing Assay Using Incucyte

For the wound healing assay, vector control or *ATP8A2* pseudogene-overexpressing cells were plated at high density, and 6 hr later, uniform scratch wounds were made using Woundmaker (Incucyte). Relative migration potential of the cells was assessed by confluence measurements at regular time intervals as indicated over the wound area.

ATP8A2 Pseudogene Overexpression Studies

The *ATP8A2* pseudogene cDNA from breast cancer cell line HCC1806 was cloned into pENTR-D-TOPO entry vector (Invitrogen) following manufacturer's instructions. Sequence-confirmed entry clones in correct orientation were recombined into Gateway pcDNA-DEST26 mammalian expression vector (Invitrogen) by LR Clonase II enzyme reaction following manufacturer's instructions. HMEC-TERT cells were transfected using Fugene 6, and polyclonal populations of cells expressing *ATP8A2* pseudogene cDNA or empty vector constructs were selected using geneticin. At the indicated times, cell numbers were measured using Coulter Counter.

Chicken Chorioallantoic Membrane Assay

Chicken chorioallantoic membrane (CAM) assay for tumor growth was carried out as follows. Fertilized eggs were incubated in a humidified incubator at 38°C for 10 days, and then CAM was dropped by drilling two holes: a small hole through the eggshell into the air sac and a second hole near the allantoic vein that penetrates the eggshell membrane but not the CAM. Subsequently, a cutoff wheel (Dremel) was used to cut a 1 cm^2 window encompassing the second hole near the allantoic vein to expose the underlying CAM. When ready, CAM was gently abraded with a sterile cotton swab to provide access to the mesenchyme, and 2×10^6 cells in 50 μl volume were implanted on top. The windows were subsequently sealed and the eggs returned to the incubator. After 7 days, extraembryonic tumors were isolated and weighed. Five to ten eggs per group were used in each experiment.

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and seven tables and can be found with this article online at doi:10.1016/j.cell.2012.04.041.

ACKNOWLEDGMENTS

The authors thank Ram-Shankar Mani, Kalpana Ramnarayanan, Joseph Mierwaza, and Sooryanarayana Varambally for technical help and Javed Siddiqui for providing samples. This work was supported, in part, by the National Institutes of Health (R01CA132874), Early Detection Research Network grant UO1 CA111275, Prostate SPORE grant P50CA69568, and the Department of Defense Era of Hope grant BC075023 (A.M.C.). A.M.C. is supported by the Doris Duke Charitable Foundation Clinical Scientist Award and the Prostate Cancer Foundation. A.M.C. is an American Cancer Society Research Professor and A. Alfred Taubman Scholar.

S.K.-S., C.K.-S., and A.M.C. conceived the experiments. S.K.-S. designed and developed the pseudogene analysis pipeline and carried out bioinformatics analyses. C.K.-S., D.R.R., N.P., X.C., and Y.-M.W. performed transcriptome sequencing. S.K.-S. and T.B. carried out ELAND mapping and provided database support. C.K.-S., V.K., and S.M.D. carried out pseudogene validations. S.S., C.K.-S., S.M.D., and I.A.A. performed the experiments. J.R.P and A.S. helped in manuscript review. S.K.-S. and R.J.L. carried out gene-pseudogene correlation analyses. S.K.-S. and M.K.I. performed ChIP-Seq analysis. C.K.-S., S.K.-S., and A.M.C. wrote the manuscript, which was reviewed by all authors.

Received: June 6, 2011
Revised: January 10, 2012
Accepted: April 26, 2012
Published: June 21, 2012

REFERENCES

Bier, A., Oviedo-Landaverde, I., Zhao, J., Mamane, Y., Kandouz, M., and Batist, G. (2009). Connexin43 pseudogene in breast cancer cells offers a novel therapeutic target. Mol. Cancer Ther. *8*, 786–793.

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell *147*, 358–369.

Dormeyer, W., van Hoof, D., Braam, S.R., Heck, A.J., Mummery, C.L., and Krijgsveld, J. (2008). Plasma membrane proteomics of human embryonic stem cells and human embryonal carcinoma cells. J. Proteome Res. *7*, 2936–2951.

Gang, L., Janecka, J.E., and Murphy, W.J. (2011). Accelerated evolution of *CES7*, a gene encoding a novel major urinary protein in the Cat family. Mol. Biol. Evol. *28*, 911–920.

Han, H., Nutiu, R., Moffat, J., and Blencowe, B.J. (2011a). SnapShot: High-throughput sequencing applications. Cell *146*, 1044.

Han, Y.J., Ma, S.F., Yourek, G., Park, Y.D., and Garcia, J.G. (2011b). A transcribed pseudogene of MYLK promotes cell proliferation. FASEB J. *25*, 2305–2312.

Harrison, P.M., Zheng, D., Zhang, Z., Carriero, N., and Gerstein, M. (2005). Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res. *33*, 2374–2383.

Kandouz, M., Bier, A., Carystinos, G.D., Alaoui-Jamali, M.A., and Batist, G. (2004). Connexin43 pseudogene is expressed in tumor cells and inhibits growth. Oncogene *23*, 4763–4770.

Karreth, F.A., Tay, Y., Perna, D., Ala, U., Tan, S.M., Rust, A.G., DeNicola, G., Webster, K.A., Weiss, D., Perez-Mancera, P.A., et al. (2011). In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. Cell *147*, 382–395.

Karro, J.E., Yan, Y., Zheng, D., Zhang, Z., Carriero, N., Cayting, P., Harrrison, P., and Gerstein, M. (2007). Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res. *35* (Database issue), D55–D60.

Kastler, S., Honold, L., Luedeke, M., Kuefer, R., Möller, P., Hoegel, J., Vogel, W., Maier, C., and Assum, G. (2010). POU5F1P1, a putative cancer susceptibility gene, is overexpressed in prostatic carcinoma. Prostate *70*, 666–674.

Katoh, M., and Katoh, M. (2003). IGSF11 gene, frequently up-regulated in intestinal-type gastric cancer, encodes adhesion molecule homologous to CXADR, FLJ22415 and ESAM. Int. J. Oncol. *23*, 525–531.

Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. Genome Res. *12*, 656–664.

Khachane, A.N., and Harrison, P.M. (2010). Mining mammalian transcript data for functional long non-coding RNAs. PLoS ONE *5*, e10316.

Khoo, C., Blanchard, R.K., Sullivan, V.K., and Cousins, R.J. (1997). Human cysteine-rich intestinal protein: cDNA cloning and expression of recombinant protein and identification in human peripheral blood mononuclear cells. Protein Expr. Purif. *9*, 379–387.

Lai, J., Lehman, M.L., Dinger, M.E., Hendy, S.C., Mercer, T.R., Seim, I., Lawrence, M.G., Mattick, J.S., Clements, J.A., and Nelson, C.C. (2010). A variant of the KLK4 gene is expressed as a cis sense-antisense chimeric transcript in prostate cancer cells. RNA *16*, 1156–1166.

Lam, H.Y., Khurana, E., Fang, G., Cayting, P., Carriero, N., Cheung, K.H., and Gerstein, M.B. (2009). Pseudofam: the pseudogene families database. Nucleic Acids Res. *37* (Database issue), D738–D743.

Lu, W., Zhou, D., Glusman, G., Utleg, A.G., White, J.T., Nelson, P.S., Vasicek, T.J., Hood, L., and Lin, B. (2006). KLK31P is a novel androgen regulated and transcribed pseudogene of kallikreins that is expressed at lower levels in prostate cancer cells than in normal prostate cells. Prostate *66*, 936–944.

Morozova, O., Hirst, M., and Marra, M.A. (2009). Applications of new sequencing technologies for transcriptome analysis. Annu. Rev. Genomics Hum. Genet. *10*, 135–151.

Munz, M., Baeuerle, P.A., and Gires, O. (2009). The emerging role of EpCAM in cancer and stem cell signaling. Cancer Res. *69*, 5627–5629.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.

Palanisamy, N., Ateeq, B., Kalyana-Sundaram, S., Pflueger, D., Ramnarayanan, K., Shankar, S., Han, B., Cao, Q., Cao, X., Suleman, K., et al. (2010). Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nat. Med. *16*, 793–798.

Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., and Pandolfi, P.P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature *465*, 1033–1038.

Pong, R.C., Lai, Y.J., Chen, H., Okegawa, T., Frenkel, E., Sagalowsky, A., and Hsieh, J.T. (2003). Epigenetic regulation of coxsackie and adenovirus receptor (CAR) gene promoter in urogenital cancer cells. Cancer Res. *63*, 8680–8686.

Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D., et al. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat. Biotechnol. *29*, 742–749.

Rieger, M.A., Ebner, R., Bell, D.R., Kiessling, A., Rohayem, J., Schmitz, M., Temme, A., Rieber, E.P., and Weigle, B. (2004). Identification of a novel mammary-restricted cytochrome P450, CYP4Z1, with overexpression in breast carcinoma. Cancer Res. *64*, 2357–2364.

Ruan, Y., Ooi, H.S., Choo, S.W., Chiu, K.P., Zhao, X.D., Srinivasan, K.G., Yao, F., Choo, C.Y., Liu, J., Ariyaratne, P., et al. (2007). Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). Genome Res. *17*, 828–838.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell *146*, 353–358.

Sasidharan, R., and Gerstein, M. (2008). Genomics: protein fossils live on as RNA. Nature *453*, 729–731.

Solovyev, V., Kosarev, P., Seledsov, I., and Vorobyev, D. (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. *7* (*Suppl 1*), S10.1–S10.12.

Sumazin, P., Yang, X., Chiu, H.S., Chung, W.J., Iyer, A., Llobet-Navas, D., Rajbhandari, P., Bansal, M., Guarnieri, P., Silva, J., and Califano, A. (2011). An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. Cell *147*, 370–381.

Tam, O.H., Aravin, A.A., Stein, P., Girard, A., Murchison, E.P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R.M., and Hannon, G.J. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature *453*, 534–538.

Tay, Y., Kats, L., Salmena, L., Weiss, D., Tan, S.M., Ala, U., Karreth, F., Poliseno, L., Provero, P., Di Cunto, F., et al. (2011). Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. Cell *147*, 344–357.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., et al. (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature *453*, 539–543.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature *453*, 1239–1243.

Yao, A., Charlab, R., and Li, P. (2006). Systematic identification of pseudogenes through whole genome expression evidence profiling. Nucleic Acids Res. *34*, 4477–4485.

Zhang, Z., Carriero, N., Zheng, D., Karro, J., Harrison, P.M., and Gerstein, M. (2006). PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics *22*, 1437–1439.

Zhang, Z., and Gerstein, M. (2004). Large-scale analysis of pseudogenes in the human genome. Curr. Opin. Genet. Dev. *14*, 328–335.

Zhang, Z.D., Frankish, A., Hunt, T., Harrow, J., and Gerstein, M. (2010). Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol. *11*, R26.

Zheng, D., and Gerstein, M.B. (2006). A computational approach for identifying pseudogenes in the ENCODE regions. Genome Biol. *7* (*Suppl 1*), S13, 1–10.

Zheng, D., Frankish, A., Baertsch, R., Kapranov, P., Reymond, A., Choo, S.W., Lu, Y., Denoeud, F., Antonarakis, S.E., Snyder, M., et al. (2007). Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res. *17*, 839–851.

Zhou, B.S., Beidler, D.R., and Cheng, Y.C. (1992). Identification of antisense RNA transcripts from a human DNA topoisomerase I pseudogene. Cancer Res. *52*, 4280–4285.

Zou, M., Baitei, E.Y., Alzahrani, A.S., Al-Mohanna, F., Farid, N.R., Meyer, B., and Shi, Y. (2009). Oncogenic activation of MAP kinase by BRAF pseudogene in thyroid tumors. Neoplasia *11*, 57–65.

NEOPLASIA
www.neoplasia.com

Volume 14 Number 8 August 2012 pp. 702–708

Gene Fusions Associated with Recurrent Amplicons Represent a Class of Passenger Aberrations in Breast Cancer[1,2]

Shanker Kalyana-Sundaram[*,†,‡], Sunita Shankar[*,†], Scott DeRoo[*,†], Matthew K. Iyer[*,§], Nallasivam Palanisamy[*,†], Arul M. Chinnaiyan[*,†,§,¶,3] and Chandan Kumar-Sinha[*,†,3]

[*]Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, MI; [†]Department of Pathology, University of Michigan, Ann Arbor, MI; [‡]Department of Environmental Biotechnology, Bharathidasan University, Tiruchirappalli, India; [§]Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI; [¶]Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, MI

Abstract

Application of high-throughput transcriptome sequencing has spurred highly sensitive detection and discovery of gene fusions in cancer, but distinguishing potentially oncogenic fusions from random, “passenger” aberrations has proven challenging. Here we examine a distinctive group of gene fusions that involve genes present in the loci of chromosomal amplifications—a class of oncogenic aberrations that are widely prevalent in breast cancers. Integrative analysis of a panel of 14 breast cancer cell lines comparing gene fusions discovered by high-throughput transcriptome sequencing and genome-wide copy number aberrations assessed by array comparative genomic hybridization, led to the identification of 77 gene fusions, of which more than 60% were localized to amplicons including 17q12, 17q23, 20q13, chr8q, and others. Many of these fusions appeared to be recurrent or involved highly expressed oncogenic drivers, frequently fused with multiple different partners, but sometimes displaying loss of functional domains. As illustrative examples of the “amplicon-associated” gene fusions, we examined here a recurrent gene fusion involving the mediator of mammalian target of rapamycin signaling, *RPS6KB1* kinase in BT-474, and the therapeutically important receptor tyrosine kinase *EGFR* in MDA-MB-468 breast cancer cell line. These gene fusions comprise a minor allelic fraction relative to the highly expressed full-length transcripts and encode chimera lacking the kinase domains, which do not impart dependence on the respective cells. Our study suggests that amplicon-associated gene fusions in breast cancer primarily represent a by-product of chromosomal amplifications, which constitutes a subset of passenger aberrations and should be factored accordingly during prioritization of gene fusion candidates.

Neoplasia (2012) 14, 702–708

Address all correspondence to: Assistant Professor Chandan Kumar-Sinha, PhD, Department of Pathology, University of Michigan Medical School, 1400 E Medical Center Dr 5316 CCGC, Ann Arbor, MI 48109-0602. E-mail: chakumar@med.umich.edu

[1] This project was supported in part by Department of Defense Breast Cancer Research Program (W81XWH-08-0110), American Association for Cancer Research Stand Up to Cancer (SU2C) award, the National Functional Genomics Center (W81XWH-11-1-0520) from Department of Defense to A.M.C., and, in part, by the US National Institutes of Health through the University of Michigan's Cancer Center Support grant 5 P30 CA46592. A.M.C. is supported by the US National Cancer Institute's Early Detection Research Network (U01 CA111275). A.M.C. is supported by the Doris Duke Charitable Foundation Clinical Scientist Award and the Burroughs Welcome Foundation Award in Clinical Translational Research. A.M.C. is an American Cancer Society Research professor and Taubman Scholar. S.D. is supported by Howard Hughes Medical Institute Medical Student Fellowship.

[2] This article refers to supplementary materials, which are designated by Tables W1 and W2 and Figures W1 to W4 and are available online at www.neoplasia.com.

[3] These authors share senior authorship.

Received 6 June 2012; Revised 9 July 2012; Accepted 10 July 2012

DOI 10.1593/neo.12914

Introduction

Chromosomal amplifications and translocations are among the most common somatic aberrations in cancers [1,2]. Gene amplification is an important mechanism for oncogene overexpression and activation. Numerous recurrent loci of chromosomal amplifications have been characterized in breast cancer, which result in gain of copy number and overexpression of oncogenes such as *ERBB2* on 17q12 (the definitive molecular aberration in 20%-30% of all breast cancers) [3,4], as well as many other oncogenic drivers including *Myc* [5], *EGFR* [6], *FGFR1* [7], *CyclinD1* [8], *RPS6KB1* [9], and others [10]. Chromosomal translocations leading to generation of gene fusions represent another prevalent mechanism for the expression of oncogenes in epithelial cancers [11]. Recently, we described the discovery and characterization of recurrent gene fusions in breast cancer involving MAST family serine threonine kinases and Notch family of transcription factors [12]. Interestingly, we also observed a large number of gene fusions, including some recurrent fusions involving known oncogenes localized at loci of chromosomal amplifications.

Here we carried out a systematic analysis of the association between gene fusions and genomic amplification by integrating RNA-Seq data with array comparative genomic hybridization (aCGH)–based whole-genome copy number profiling from a panel of breast cancer cell lines. We examined a set of "amplicon-associated gene fusions" that refer to all the fusions where one or both gene partners are localized to a site of chromosomal amplification. Specifically, we assessed the functional relevance of two amplicon-associated fusion genes involving oncogenic kinases *EGFR* and *RPS6KB1* in the context of prioritizing fusion candidates important in tumorigenesis. Our results suggest that recurrent gene fusions localized to recurrent amplicons, displaying allelic imbalance between the fusion partners, may represent an epiphenomenon of genomic amplification cycles not essential for cancer development.

Materials and Methods

Gene Fusion Data Set

Chimeric transcript candidates were primarily obtained from paired-end transcriptome sequencing of breast cancer from a total of more than 49 cell lines and 40 tissue samples described previously [12]. aCGH data were generated using Agilent Human Genome 244A CGH Microarrays (Agilent Technologies, Santa Clara, CA) according to the manufacturer's instructions, and data were analyzed using CGH Analytics (Agilent Technologies). Copy number alterations were assessed using ADM-2, with the threshold a setting of 6.0 and a bin size of 10.

RNA Isolation and Complementary DNA Synthesis

Total RNA was isolated using TRIzol and RNeasy Kit (Invitrogen, Carlsbad, CA) with DNase I digestion according to the manufacturer's instructions. RNA integrity was verified on an Agilent Bioanalyzer 2100 (Agilent Technologies). Complementary DNA was synthesized from total RNA using Superscript III (Invitrogen) and random primers (Invitrogen).

Quantitative Real-time Polymerase Chain Reaction

Primers for validation of candidate gene fusions were designed using the National Center for Biotechnology Information Primer Blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast/), with primer pairs spanning exon junctions amplifying 70- to 110-bp products for every chimera tested. Quantitative polymerase chain reaction (QPCR) was performed using SYBR Green MasterMix (Applied Biosystems, Carlsbad, CA) on an Applied Biosystems StepOne Plus Real-Time PCR System. All oligonucleotide primers were obtained from Integrated DNA Technologies and are listed in Table W1. *GAPDH* was used as endogenous control. All assays were performed twice, and results were plotted as average fold change relative to *GAPDH*.

Cell Proliferation Assays

Cells were transfected with small interfering RNAs (siRNAs) using Oligofectamine reagent (Life Sciences, Carlsbad, CA), and 3 days after transfection, the cells were plated for proliferation assays. At the indicated times, cell numbers were counted using Coulter Counter (Indianapolis, IN).

Western Blot

Cell pellets were sonicated in NP-40 lysis buffer (50 mM Tris-HCl, 1% NP-40, pH 7.4; Sigma, St. Louis, MO), complete protease inhibitor mixture (Roche, Indianapolis, IN), and phosphatase inhibitor (EMD Bioscience, San Diego, CA). Immunoblot analysis was carried out using antibodies for *ERBB2* (MS-730-PABX; Thermo Scientific, Fremont, CA) and *RPS6KB1* (2708S; Cell Signaling, Danvers, MA). Human β-actin antibody (Sigma, St. Louis, MO) was used as a loading control.

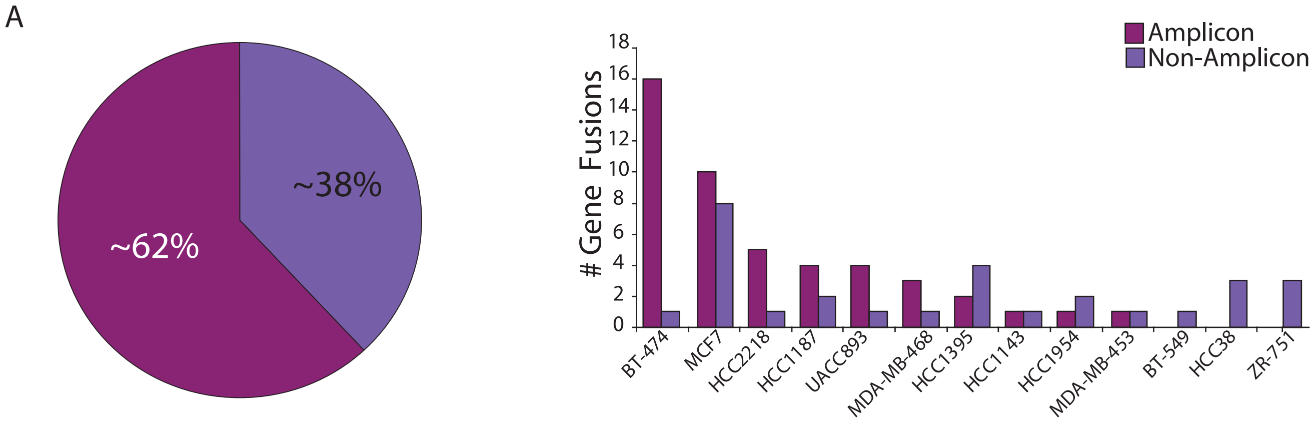
Knockdown Assays

Short hairpin RNAs (shRNAs; Table W1) were transduced in presence of 1 μg/ml polybrene. All siRNA transfections were performed using Oligofectamine reagent (Life Sciences). For siRNA knockdown experiments, multiple custom siRNA sequences targeting the *ARID1A-MAST2* fusion (Thermo, Lafayette, CO) were used [12].

Results

Paired-end transcriptome sequencing of breast cancer cell lines and tissues led to the identification of an average of more than four gene fusions per breast cancer sample [12]. Interestingly, we observed that some of the cell lines with the largest number of gene fusions also harbored many well-known chromosomal amplifications, prompting us to examine a likely association between genomic amplifications and gene fusions. To assess copy number alterations at the chromosomal coordinates of the fusion genes, we analyzed aCGH (244K Agilent array) data in a set of 14 cell lines (Table W2) and observed that as many as 62% of the total number of fusions were associated with regions of amplifications (Figure 1*A*). The genes involved in fusions were found to be significantly associated with their genomic amplification status based on Fisher exact *t* test (*P* < .0004), in four of six cell lines with the maximum number of fusions, including BT-474, MCF7, HCC2218, and UACC893 (Figure 1*B*).

Examining the distribution of fusion genes in individual samples revealed that a majority of the gene fusions were associated with 17q12 amplicon harboring *ERBB2* and 17q23 amplicon that includes genes such as *BCAS3*, *RPS6KB1*, and *TMEM49*, 20q13 amplicon with *BCAS4* and the chr8q amplicon commonly found amplified in breast cancer (Table W2 and Figures 2 and W1). Interestingly, the breast cancer cell line BT-474 that harbors both the chr17 amplicons and the chr20 amplicon and MCF7 with prominent amplifications in chr17, chr20, and chr8q showed the maximum number of gene fusions observed in a sample, accounting for as many as 26 gene fusions associated with amplicons compared against only 9 in unamplified loci (Figures 1 and 2 and Table W2).



B

Sample	Fusion Partners	Number of Amplifications	Total Measureable Genes	Amplified Fusion Partners	Not-amplified Fusion Partners	Amplified without Fusion Partner	Unaffected Genes	p-value
MCF7	27	426	17691	11	16	415	17249	1.28E-11
HCC2218	9	446	17691	7	2	439	17243	2.13E-10
BT-474	31	5648	17691	25	6	5623	12038	3.19E-08
UACC-893	8	2902	17691	6	2	2896	14787	0.0004016
HCC1187	10	780	17691	2	8	778	16903	0.06908
HCC1395	12	2111	17691	1	11	2110	15569	0.783

Figure 1. Distribution of gene fusions across breast cancer cell lines. (A) Pie chart representation of the relative proportion of gene fusions associated with loci of genomic amplifications compared to unamplified loci (left) and bar graph representation of the relative distribution of gene fusions across different breast cancer cell lines (right). (B) Table summarizing the statistical significance of association between gene fusions and chromosomal amplifications in breast cancer cell lines with the highest number of gene fusions in A (using Fisher exact *t* test, sorted by *P* value).

In the backdrop of a large number of somatic aberrations seen in cancers, any "recurrent" events observed across samples are generally regarded as potentially "driving" tumorigenesis. Interestingly, among the more than 380 gene fusions reported in our compendium of breast cancer fusions [12], as many as 62 genes were found to be recurrent partners (appear at least twice). Among these, whereas the *MAST* and *Notch* fusions were shown to be functionally recurrent and potentially driving aberrations in up to 5% to 7% of breast cancers, 33 of other recurrent gene fusions were found to be associated with known frequent amplicons, including *ERBB2*, *BCAS3/4*, and chr8q. Among these, three fusions each involved the ikaros family zinc finger protein 3 transcription factor (*IKZF3* on chr17q12 amplicon) and breast carcinoma amplified sequence 3 (*BCAS3* on chr17q23 amplicon) as 3′ partners—all with different 5′ partners. Similarly, tripartite motif containing 37 (*TRIM37* on chr17q23) was a common 5′ partner in three distinct gene fusions with different 3′ partners (Table W2). To further expand our integrative analysis of copy number aberrations and gene fusions, next we used the breast cancer aCGH data [13,14] and observed gene fusion–associated amplicons in MCF7, BT-474, and MDA-MB-468, HCC-1187 as seen in our data as well as in an additional panel of cell lines, including ZR-75-30, SUM190, MDA-MB-361, HCC-1428, and HCC-1569 (Figure W2). Clearly, apart from triggering overexpression of constituent genes, our observations strongly suggest that the loci of chromosomal amplifications also serve as "hotspots" for the generation of recurrent gene fusions.

Next, to assess whether amplicon-associated gene fusions impart oncogenic phenotypes on the cells, we examined the open reading frames (ORFs), functional domains/motifs, and conservation of fusion architecture across different samples. Among recurrent fusion candidates within amplicons, we focused on known cancer-associated partner genes such as kinases, oncogenes, tumor suppressors, or known fusion partners in the Mitelman Database of chromosomal aberrations in cancer [15] and observed several functionally plausible gene fusions. Here we describe our observations with two specific examples of gene fusions involving oncogenic kinases.

The triple-negative breast cancer cell line MDA-MB-468 is known to show an overexpression of epidermal growth factor receptor (EGFR) [16]. In our transcriptome sequencing compendium of 89 breast cancer cell lines and tissues, the highest expression of *EGFR* is observed in MDA-MB-468 (Figure 3*A*), potentially resulting from a focal amplification at chr7p12 (Figure 2). In addition, we detected an *EGFR* fusion transcript (*EGFR-POLD1*) in this cell line, encoding the N-terminal portion of EGFR, completely devoid of the tyrosine kinase domain (Figure 3*A*, *inset*). However, the uniform read-coverage observed across the full length of the *EGFR* transcript in this sample (Figure 3*B*), precluded the existence of any exon imbalance, suggesting that even as the kinase domain is lost in the fusion, the full-length EGFR protein is expressed in this cell line. Further, we observed a remarkable mismatch between the copy numbers of *EGFR* and its fusion partner *POLD1* (Figure 3*C*) that supports a predominant expression of full-length *EGFR* compared with the *EGFR-POLD1* chimera. This is unlike the observation in case of *MAST* kinase fusions in breast cancer characterized in our previous study [12], in which case a marked exon imbalance in coverage was observed (Figure W3). Considering that the

MDA-MB-468 harbors both *MAST2* and *EGFR* fusions, we were intrigued to assess its relative "dependence" on both the kinases. Surprisingly, a profound reduction in cell proliferation was observed on siRNA knockdown of *MAST2*, whereas *EGFR* knockdown showed little effect (Figure 3*D*). Next, testing the possibility of *EGFR* amplicon potentially cooperating with *MAST2*, we found that the effect of combined knockdown of *EGFR* and *MAST2* was comparable with that of *MAST2* knockdown alone (Figure 3*D*), further suggesting that *EGFR* amplification does not signify a driver aberration. In this context, the EGFR fusion transcript that represents a miniscule fraction of overall EGFR expression and encodes only the N-terminal portion lacking the kinase domain was reckoned to be inconsequential.

Next, we looked at recurrent gene fusions involving oncogenic serine threonine kinase ribosomal protein S6 kinase on chr17q23 frequently amplified in breast cancers [17–20] identified in BT-474 (*RPS6KB1-SNF8*) and MCF7 (*RPS6KB1-VMP1*). Both of these cell lines harbor amplifications at the *RPS6KB1* locus and express the highest levels of *RPS6KB1* among all the samples examined (Figure 4*A*). Both the chimeric transcripts retain only the first exon of *RPS6KB1* and the respective open reading frames show a complete loss of the kinase domain (Figure 4*A*, *inset*). We also observed an even read coverage across the *RPS6KB1* transcript in both fusion-positive cell lines, similar to a representative benign mammary epithelial cell line, albeit at a much higher level, indicating that full-length *RPS6KB1* protein is encoded in these samples (Figures 4*B* and W4*A*). Further, the difference between the copy number observed between the fusion partners in both the *RPS6KB1* fusions (Figures 4*C* and W4*B*) indicates an allelic imbalance between the full-length and the putative fusion genes. Next, considering that BT-474 is an *ERBB2*-positive cell line, we tested potential dependence of these cells on the RPS6KB1 protein. Surprisingly,

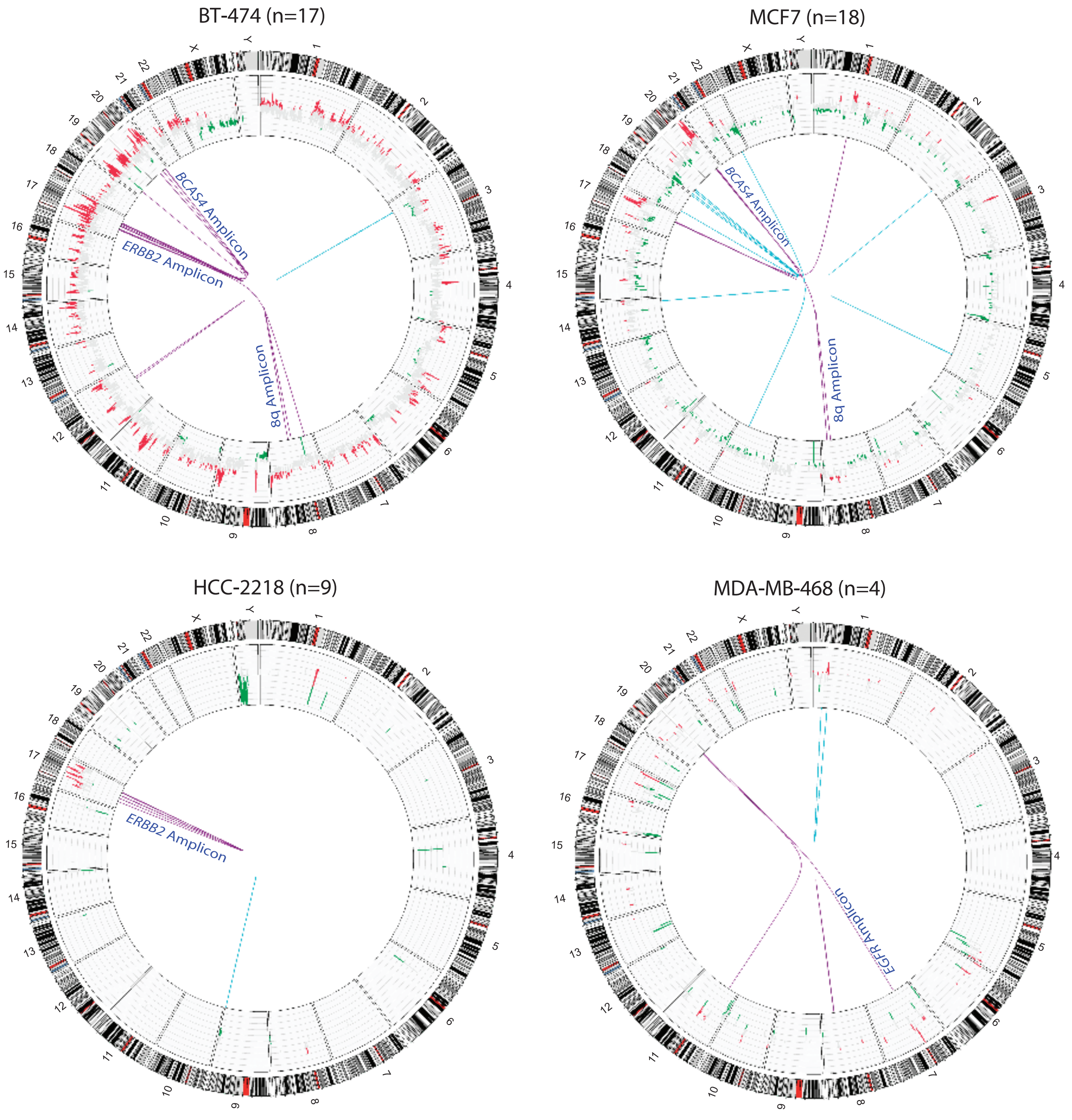


Figure 2. Graphical representation of integrative analysis of gene fusions with copy number analysis. Circos plots of the genome-wide distribution of gene fusions along with status of copy number alterations. Red and green peaks represent amplifications and deletions; purple and cyan lines represent the fusions associated with amplicons and nonamplicons, respectively. "*n*" refers to the total number of fusions identified.

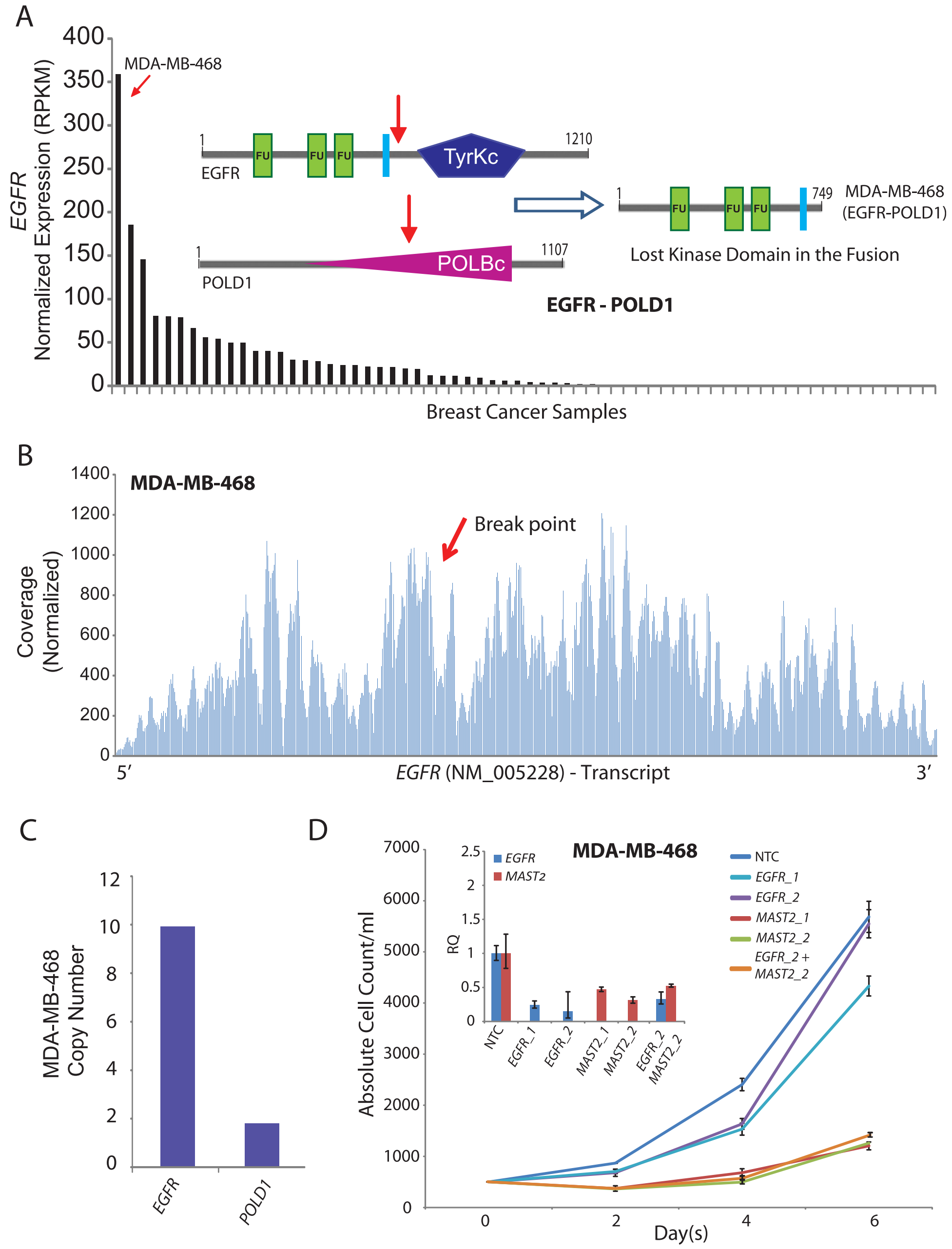


Figure 3. (A) Normalized expression (RPKM) of *EGFR* in descending order of expression in a panel of breast cancer samples obtained from RNA-Seq. Schematic representation of wild-type EGFR and POLD1 proteins with putative breakpoints indicated by red arrows and the domain structure of the putative fusion protein (inset). (B) Plot of normalized coverage of *EGFR* transcript in MDA-MB-468 cell line showing the location of the breakpoint (indicated by red arrow). (C) Bar graph representing the copy number of *EGFR* and *POLD1* in MDA-MB-468. (D) Proliferation assay showing absolute cell count (*y* axis) over a time course (*x* axis) after knockdown with *EGFR* and/or *MAST2* siRNAs in MDA-MB-468. QPCR assessment of knockdown efficiencies relative to nontargeted control (NTC; inset).

similar to our observations with *EGFR* knockdown in MDA-MB-468 cells, here we observed only a small effect on cell proliferation after shRNA knockdown of *RPS6KB1*, in dramatic contrast to the effect of *ERBB2* knockdown (Figure 4*D*). Notably, the shRNA knockdown of RPS6KB1 led to a significant depletion of the full-length protein yet it did not affect cell proliferation compared with ERBB2 protein depletion (Figure 4*D*, *inset*). Therefore, BT-474 cells do not display a dependence on RPS6KB1 protein, and considering that the RPS6KB1 fusion product is completely devoid of all functional domains of RPS6KB1, including the kinase domain, this fusion also likely represents a passenger event.

Discussion

In our systematic search for gene fusions in breast cancer using high-throughput transcriptome sequencing, we observed a notably large number of fusion genes associated with many well characterized recurrent amplicons, including 17q12, 17q23, 20q13, and 8q, among others. Amplicon-associated gene fusions were found to involve complex and cryptic rearrangements, involving one or both partners within the amplicon site, with the chimeric transcript expression apparently concealed in the backdrop of highly expressed wild-type genes. The gene fusions considered here include only “expressed” chimeric transcripts derived from known/annotated fusion partners. Chromosomal rearrangements that do not express chimeric transcripts or that involve unannotated fusion partners are excluded from this analysis. This likely accounts for the variability observed in the number of gene fusions scored across multiple samples with known amplicons. Because many of the fusions at the amplicons appeared to be recurrent, although frequently fused with multiple different partners, it led us to examine whether the recurrence was incidentally associated with recurrent amplicons or signified functionally important aberrations.

MDA-MB-468 represents a prototype triple-negative breast cancer cell line with a "basal-like" gene expression profile that shows an overexpression of the oncogenic kinase *EGFR* due to a focal amplification at chr7p12. Here we discovered a chimeric transcript involving *EGFR*. However, careful examination of this transcript revealed that the fusion encodes N-terminal *EGFR* protein, without the kinase domain. Transcriptome sequencing did not show evidence of fusion-associated exon imbalance in *EGFR* expression, suggesting that full-length *EGFR* is expressed in this cell line. In addition, the significantly higher genomic copy number of *EGFR* compared to its fusion partner *POLD1* suggests that a minor allelic fraction of the *EGFR* is involved in fusion with *POLD1*, whereas other amplified copies of the gene

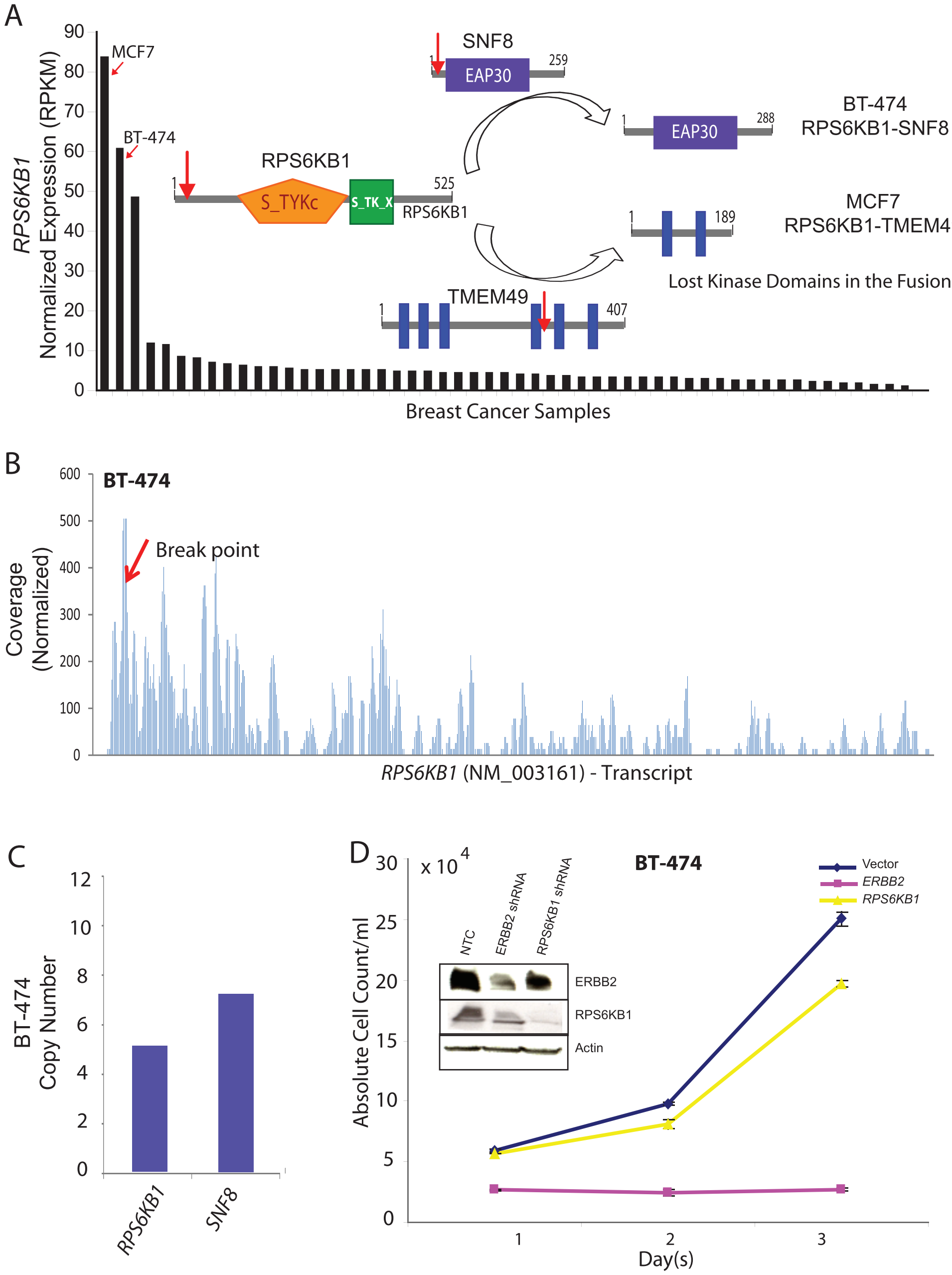


Figure 4. (A) Normalized expression (RPKM) of *RPS6KB1* in descending order of expression in a panel of breast cancer samples obtained from RNA-Seq. Schematic representation of wild-type RPS6KB1, TMEM49, and SNF8 proteins with putative breakpoints indicated by red arrows and the domain structure of the putative fusion proteins in BT-474 and MCF7 (inset). (B) Plot of normalized coverage of *RPS6KB1* transcript in BT-474 cell line showing the location of the breakpoint (indicated by red arrow). (C) Bar graph representing the copy number of *RPS6KB1* and *SNF8* in BT-474 (D) Proliferation assay showing absolute cell count (*y* axis) over a time course (*x* axis) after knockdown with *RPS6KB1* and/or *ERBB2* shRNAs in BT-474. Western blot assessment of the knockdown efficiency relative to nontargeted control (NTC). Actin was used as a loading control (inset).

express the full-length molecule. Technically, the detection and monitoring of the *EGFR* fusion transcript in the backdrop of extremely high levels of wild-type *EGFR* transcript is challenging; therefore, we chose to assess the dependency imparted by full-length *EGFR*. Interestingly, the knockdown of *EGFR* had only a slight effect on the proliferation of MDA-MB-468 cells, whereas a profound reduction in cell proliferation was observed on the knockdown the fusion gene *MAST2.* Combined knockdown of *MAST2* and *EGFR* produced the same effect as that by MAST2 alone, further calling into question the credentials of *EGFR* as a driver aberration in MDA-MB-468 cells. Interestingly, MDA-MB-468 is known to be insensitive to EGFR inhibitors like erlotinib [21] and gefitinib [22].

Similarly, the recurrent gene fusions involving *RPS6KB1* retain only the first exon, and the chimeric ORFs show a complete loss of the kinase domain in breast cancer cell lines BT-474 and MCF7. Similar to the *EGFR* fusion, DNA copy number analysis and RNA-Seq data provided the evidence that full-length RPS6KB1 protein is encoded in both these cell lines. Notably, both BT-474 and MCF7 have been shown to express high levels of full-length RPS6KB1 protein [23], suggesting that these cells exhibit elevated activity of RPS6KB1 as a result of amplification, independent of the fusion. Again, similar to *EGFR* knockdown in MDA-MB-468, *RPS6KB1* knockdown in BT-474 (an ERBB2-positive cell line) showed an insignificant effect on cell proliferation compared to *ERBB2* knockdown. Interestingly, in a previous study, knockdown of *RPS6KB1* was found to have no effect on apoptosis in both BT-474 and MCF7 breast cancer cells [24].

In the light of our observations, we surmise that repeated breaks and rejoining of chromosomes during chromosomal amplifications led to the generation of amplicon-associated gene fusions. Loci of recurrent genomic amplifications thus engender “pseudo” recurrent gene fusions that may largely represent passenger aberrations involving random breakpoints. The two cell lines with established drivers—*ERBB2* in BT-474 and *MAST2* in MDA-MB-468—made it possible for us to assess the relative importance of amplicon fusions involving *RPS6KB1* and *EGFR*, respectively. In cases where a driver is not clearly apparent, a more careful examination of all plausible fusion candidates will be required. Importantly, even as our study primarily pertains to breast cancers based on available data and a well-documented preponderance of copy number aberrations in breast cancers [10], we expect the association between amplicons and gene fusions to be consistent in other cancers as well. We argue here for a measure of caution in considering the functional implications of recurrent gene fusions associated with amplifications because these may be simply a result of massive chromosomal upheaval at the amplicons, not representing clonally selected oncogenic events.

Acknowledgments

The authors thank Vishal Kothari and Nikita Consul for technical help.

References

[1] Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, and Stratton MR (2004). A census of human cancer genes. *Nat Rev Cancer* **4**, 177–183.

[2] Santarius T, Shipley J, Brewer D, Stratton MR, and Cooper CS (2010). A census of amplified and overexpressed human cancer genes. *Nat Rev Cancer* **10**, 59–64.

[3] Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, and McGuire WL (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/*neu* oncogene. *Science* **235**, 177–182.

[4] Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, et al. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N Engl J Med* **344**, 783–792.

[5] Deming SL, Nass SJ, Dickson RB, and Trock BJ (2000). C-*myc* amplification in breast cancer: a meta-analysis of its occurrence and prognostic relevance. *Br J Cancer* **83**, 1688–1695.

[6] Bhargava R, Gerald WL, Li AR, Pan Q, Lal P, Ladanyi M, and Chen B (2005). EGFR gene amplification in breast cancer: correlation with epidermal growth factor receptor mRNA and protein expression and HER-2 status and absence of EGFR-activating mutations. *Mod Pathol* **18**, 1027–1033.

[7] Elbauomy Elsheikh S, Green AR, Lambros MB, Turner NC, Grainge MJ, Powe D, Ellis IO, and Reis-Filho JS (2007). FGFR1 amplification in breast carcinomas: a chromogenic *in situ* hybridisation analysis. *Breast Cancer Res* **9**, R23.

[8] Elsheikh S, Green AR, Aleskandarany MA, Grainge M, Paish CE, Lambros MB, Reis-Filho JS, and Ellis IO (2008). CCND1 amplification and cyclin D1 expression in breast cancer and their relation with proteomic subgroups and patient outcome. *Breast Cancer Res Treat* **109**, 325–335.

[9] Inaki K, Hillmer AM, Ukil L, Yao F, Woo XY, Vardy LA, Zawack KF, Lee CW, Ariyaratne PN, Chan YS, et al. (2011). Transcriptional consequences of genomic structural aberrations in breast cancer. *Genome Res* **21**, 676–687.

[10] Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352.

[11] Chinnaiyan AM and Palanisamy N (2010). Chromosomal aberrations in solid tumors. *Prog Mol Biol Transl Sci* **95**, 55–94.

[12] Robinson DR, Kalyana-Sundaram S, Wu YM, Shankar S, Cao X, Ateeq B, Asangani IA, Iyer M, Maher CA, Grasso CS, et al. (2011). Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nat Med* **17**, 1646–1651.

[13] Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, and Brown PO (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci USA* **99**, 12963–12968.

[14] Greshock J, Naylor TL, Margolin A, Diskin S, Cleaver SH, Futreal PA, deJong PJ, Zhao S, Liebman M, and Weber BL (2004). 1-Mb resolution array-based comparative genomic hybridization using a BAC clone set optimized for cancer gene analysis. *Genome Res* **14**, 179–187.

[15] Mitelman F, Johansson B, and Mertens F (2010). *Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer*. Cancer Genome Anatomy Project. Available at: http://cgap.nci.nih.gov/Chromosomes/Mitelman. Accessed March 2012.

[16] Hyatt DC and Ceresa BP (2008). Cellular localization of the activated EGFR determines its effect on cell growth in MDA-MB-468 cells. *Exp Cell Res* **314**, 3415–3425.

[17] Barlund M, Monni O, Kononen J, Cornelison R, Torhorst J, Sauter G, Kallioniemi O-P, and Kallioniemi A (2000). Multiple genes at 17q23 undergo amplification and overexpression in breast cancer. *Cancer Res* **60**, 5340–5344.

[18] Couch FJ, Wang XY, Wu GJ, Qian J, Jenkins RB, and James CD (1999). Localization of PS6K to chromosomal region 17q23 and determination of its amplification in breast cancer. *Cancer Res* **59**, 1408–1411.

[19] Monni O, Barlund M, Mousses S, Kononen J, Sauter G, Heiskanen M, Paavola P, Avela K, Chen Y, Bittner ML, et al. (2001). Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. *Proc Natl Acad Sci USA* **98**, 5711–5716.

[20] Sinclair CS, Rowley M, Naderi A, and Couch FJ (2003). The 17q23 amplicon and breast cancer. *Breast Cancer Res Treat* **78**, 313–322.

[21] Bartholomeusz C, Yamasaki F, Saso H, Kurisu K, Hortobagyi GN, and Ueno NT (2011). Gemcitabine overcomes erlotinib resistance in EGFR-overexpressing cancer cells through downregulation of Akt. *J Cancer* **2**, 435–442.

[22] Maiello MR, D’Alessio A, De Luca A, Carotenuto A, Rachiglio AM, Napolitano M, Cito L, Guzzo A, and Normanno N (2007). AZD3409 inhibits the growth of breast cancer cells with intrinsic resistance to the EGFR tyrosine kinase inhibitor gefitinib. *Breast Cancer Res Treat* **102**, 275–282.

[23] Yamnik RL, Digilova A, Davis DC, Brodt ZN, Murphy CJ, and Holz MK (2009). S6 kinase 1 regulates estrogen receptor alpha in control of breast cancer cell proliferation. *J Biol Chem* **284**, 6361–6369.

[24] Heinonen H, Nieminen A, Saarela M, Kallioniemi A, Klefstrom J, Hautaniemi S, and Monni O (2008). Deciphering downstream gene targets of PI3K/mTOR/p70S6K pathway in breast cancer. *BMC Genomics* **9**, 348.

Table W1. Primer Sequences and siRNA/shRNA Clone Details.

Gene Symbol	Clone ID
EGFR	LU-003114-00-0002
ERBB2	SHCLNV-NM_004448
RPS6KB1	SHCLNV-NM_003161

Primer	Sequence
EGFR-f1	GGGCCAGGTCTTGAAGGCTGT
EGFR-r1	ATCCCCAGGGCCACCACCAG
EGFR-f2	ACACCCTGGTCTGGAAGTACGCA
EGFR-r2	AGTGGGAGACTAAAGTCAGACAGTGAA
EGFR-f3	CCGAGGCAGGGAATGCGTGG
EGFR-r3	TGGCCTGAGGCAGGCACTCT
ERBB2-f1	TGCGCAGGCAGTGATGAGAGT
ERBB2-r1	TCTCGGGACTGGCAGGGAGC
ERBB2-f2	TCCTCCTCGCCCTCTTGCCC
ERBB2-r2	TCTCGGGACTGGCAGGGAGC
RPS6KB1-f1	TGCTGACTGGAGCACCCCCA
RPS6KB1-r1	GCTTCTTGTGTGAGGTAGGGAGGC
GAPDH-f1	GGCTGAGAACGGGAAGCTTGTCA
GAPDH-r1	TCTCCATGGTGGTGAAGACGCCA
MAST2_f1	GAAGTGAGTGAGGATGGCTGCCTT
MAST2_r1	GAGCCGCTCCATGCTGCTGTAC
MAST2_f2	ATTGAGGGCCATGGGGCATCT
MAST2_r2	CCCCATAGGCGCCATTGCTGATG

Table W2. List of Gene Fusions Identified in 14 Breast Cancer Cell Lines, along with Their Copy Number Status.

Sample Name	5′ Gene	3′ Gene	Type	Sequencing Platform	No. Reads	Validation	Chromosomal Location		aCGH Data (5′ and 3′)				Amplicon Status
						Fusion QPCR	5′ Gene	3′ Gene	No. Probe	Average Log Ratio	No. Probe	Average Log Ratio	
BT-474	*RPS6KB1*	SNF8	Intra	GA II	92	Y	chr17:55325224-55382568	chr17:44362457-44377153	5	2.890	2	3.557	Yes
BT-474	*STX16*	RAE1	Intra	GA II	79	Y	chr20:56659733-56687988	chr20:55360024-55386926	4	2.910	4	2.910	Yes
BT-474	*ZMYND8*	CEP250	Intra	GA II	77	Y	chr20:45271787-45418881	chr20:33506636-33563217	15	3.650	5	1.876	Yes
BT-474	*TRPC4AP*	MRPL45	Inter	GA II	30		chr20:33053867-33144279	chr17:33706516-33732628	11	3.290	4	3.452	Yes
BT-474	*MED1*	STXBP4	Intra	GA II	28	Y	chr17:34814063-34861053	chr17:50401124-50596448	4	4.029	21	2.507	Yes
BT-474	*TOB1*	AP1GBP1	Intra	GA II	16		chr17:46294585-46296412	chr17:32949013-33043559	1	2.787	10	2.556	Yes
BT-474	*ACACA*	STAC2	Intra	GA II	15		chr17:32516039-32841015	chr17:34620314-34635566	35	2.556	3	4.029	Yes
BT-474	*MED13*	BCAS3	Intra	GA II	13	Y	chr17:57374747-57497425	chr17:56109953-56824981	13	1.012	73	1.934	Yes
BT-474	*VAPB*	IKZF3	Inter	GA II	13	Y	chr20:56397580-56459562	chr17:35174724-35273967	7	3.404	10	3.701	Yes
BT-474	*RAB22A*	MYO9B	Inter	GA II	9	Y	chr20:56318176-56375969	chr19:17047590-17185104	6	3.404	13	2.122	Yes
BT-474	*GLB1*	CMTM7	Intra	GA II	7		chr3:33013103-33113698	chr3:32408166-32471337	11	–0.425	6	0.428	
BT-474	*NCOA2*	ZNF704	Intra	GA II	7	Y	chr8:71186820-71478574	chr8:81703240-81949571	35	0.916	26	0.640	Yes
BT-474	*BCAS3*	MED13	Intra	GA II	6		chr17:56109953-56824981	chr17:57374747-57497425	73	1.934	13	1.012	Yes
BT-474	*PIP4K2B*	RAD51C	Intra	GA II	6		chr17:34175469-34209684	chr17:54124961-54166691	6	4.813	5	1.700	Yes
BT-474	*PPP1R12A*	MGAT4C	Intra	GA II	6		chr12:78691473-78853366	chr12:84897167-85756812	19	1.218	90	–0.397	Yes
BT-474	*STARD3*	DOCK5	Inter	GA II	6		chr17:35046858-35073980	chr8:25098203-25326536	5	4.821	27	0.076	Yes
BT-474	*TRIM37*	MYO19	Intra	GA II	6		chr17:54414781-54539048	chr17:31925711-31965418	14	2.244	6	2.344	Yes
BT-483	*SMARCB1*	MARK3	Inter	GA II	7	Y	chr22:22459149-22506705	chr14:102921453-103039919	8	1.170	17	0.381	Yes
BT-549	*CLTC*	TMEM49	Intra	GA II	18	Y	chr17:55051831-55129099	chr17:55139644-55272734	9	–0.283	18	–1.185	
HCC1143	*C18orf45*	HM13	Inter	GA II	25	Y	chr18:19129977-19271923	chr20:29565901-29591257	18	1.280	2	1.403	Yes
HCC1143	*C2ORF48*	RRM2	Intra	GA II	23	Y	chr2:10198959-10269307	chr2:10180145-10188997	8	0.134	2	0.134	
HCC1187	*PUM1*	TRERF1	Inter	GA II	38	Y	chr1:31176939-31311151	chr6:42300646-42527761	14	1.648	27	0.336	Yes
HCC1187	*SEC22B*	NOTCH2	Intra	GA II	30	Y	chr1:143807763-143828279	chr1:120255698-120413799	2	1.557	11	0.253	Yes
HCC1187	*CTAGE5*	SIP1	Intra	GA II	15		chr14:38806079-38890148	chr14:38653238-38675928	9	0.940	4	0.235	Yes
HCC1187	*MCPH1*	AGPAT5	Intra	GA II	11		chr8:6251520-6488548	chr8:6553285-6606429	29	0.495	5	0.738	
HCC1187	*KLK5*	CDH23	Inter	GA II	5		chr19:56138370-56148156	chr10:73225533-73245710	3	0.888	1	0.953	Yes
HCC1187	*BC041478*	EXOSC10	Inter	GA II	3		chr19:42434668-42446354	chr1:11049262-11082525	1	0.816	4	0.156	
HCC1395	*EIF3K*	CYP39A1	Inter	GA II	13	Y	chr19:43801561-43819435	chr6:46625403-46728482	2	0.852	11	0.611	
HCC1395	*HNRNPUL2*	AHNAK	Intra	GA II	13	Y	chr11:62238795-62251397	chr11:62039949-62070908	2	0.629	5	1.172	Yes
HCC1395	*RAB7A*	LRCH3	Intra	GA II	6		chr3:129927668-130016331	chr3:199002541-199082853	10	0.755	11	–0.615	
HCC1395	*ERO1L*	FERMT2	Intra	GA II	5		chr14:52178354-52232169	chr14:52395955-52487565	7	0.934	14	0.934	Yes
HCC1395	*FOSL2*	BRE	Intra	GA II	5		chr2:28469282-28491020	chr2:27966985-28415271	3	0.480	51	0.480	
HCC1395	*BCAR3*	ABCA4	Intra	GA II	4		chr1:93799936-93919973	chr1:94230981-94359293	13	0.849	13	0.849	
HCC1954	*C6orf106*	SPDEF	Intra	GA II	24	Y	chr6:34663048-34772603	chr6:34613557-34632069	13	0.036	3	0.374	
HCC1954	*INTS1*	PRKAR1B	Intra	GA II	22	Y	chr7:1476438-1510544	chr7:555359-718687	4	1.034			Yes
HCC1954	*GALNT7*	ORC4L	Inter	GA II	9		chr4:174326478-174481693	chr2:148408201-148494933	15	0.409	7	0.504	
HCC2218	*SEC16A*	NOTCH1	Intra	GA II	14	Y	chr9:138454368-138497328	chr9:138508716-138560059	6	0.000	7	–0.967	
HCC2218	*POLDIP2*	BRIP1	Intra	GA II	8		chr17:23697785-23708730	chr17:57111328-57295702	3	1.113	19	3.925	Yes
HCC2218	*INTS2*	ZNF652	Intra	GA II	7		chr17:57297509-57360159	chr17:44721566-44794834	9	3.925	6	2.649	Yes
HCC2218	*INTS2*	TMEM49	Intra	GA II	5		chr17:57297509-57360159	chr17:55139644-55272734	9	3.925	18	3.202	Yes
HCC2218	*LRRC59*	NEUROD2	Intra	GA II	5		chr17:45813592-45829913	chr17:35013546-35017701	3	2.649	1	3.451	Yes
HCC2218	*PERLD1*	PPM1D	Intra	GA II	4	Y	chr17:35082579-35097833	chr17:56032335-56096818	2	3.451	7	3.340	Yes
MCF7	*BCAS4*	BCAS3	Inter	GA II	2788		chr20:48844873-48927121	chr17:56109953-56824981	7	2.107	73	2.653	Yes
MCF7	*ARFGEF2*	SULF2	Intra	GA II	305	Y	chr20:46971681-47086637	chr20:45719556-45848215	11	0.823	13	3.398	Yes
MCF7	*RPS6KB1*	TMEM49	Intra	GA II	78	Y	chr17:55325224-55382568	chr17:55139644-55272734	5	3.412	18	2.197	Yes
MCF7	*STK11*	MIDN	Intra	GA II	25		chr19:1156797-1179434	chr19:1199551-1210142	4	–1.367	2	–0.279	
MCF7	*PAPOLA*	AK7	Intra	GA II	16	Y	chr14:96038472-96103201	chr14:95928200-96024865	7	0.343	13	0.343	
MCF7	*AHCYL1*	RAD51C	Inter	GA II	12	Y	chr1:110328830-110367887	chr17:54124961-54166691	4	–0.063	5	2.788	Yes
MCF7	*EIF3H*	FAM65C	Inter	GA II	11		chr8:117726235-117837243	chr20:48636052-48686833	12	0.456	5	1.554	Yes
MCF7	*BC017255*	TMEM49	Intra	GA II	10		chr17:54538741-54550409	chr17:55139644-55272734	1	3.515	18	2.197	Yes
MCF7	*ADAMTS19*	SLC27A6	Intra	GA II	9		chr5:128824001-129102275	chr5:128329108-128397234	30	0.051	8	0.051	
MCF7	*ARHGAP19*	DRG1	Inter	GA II	8	Y	chr10:98971919-99042403	chr22:30125538-30160172	8	0.387	5	–0.420	

Table W2. (*continued*)

Sample Name	5′ Gene	3′ Gene	Type	Sequencing Platform	No. Reads	Validation	Chromosomal Location		aCGH Data (5′ and 3′)				Amplicon Status
						Fusion QPCR	5′ Gene	3′ Gene	No. Probe	Average Log Ratio	No. Probe	Average Log Ratio	
MCF7	*MYO9B*	FCHO1	Intra	GA II	8	Y	chr19:17047590-17185104	chr19:17719526-17760377	13	−1.126	4	−0.529	
MCF7	*HSPE1*	PREI3	Intra	GA II	6	Y	chr2:198072965-198076432	chr2:198089016-198125760	1	−0.361	4	−0.361	
MCF7	*PARD6G*	C18ORF1	Intra	GA II	6		chr18:76016105-76106388	chr18:13601664-13642753	10	−0.674	5	−0.407	
MCF7	*TRIM37*	TMEM49	Intra	GA II	6	Y	chr17:54414781-54539048	chr17:55139644-55272734	14	3.515	18	2.197	Yes
MCF7	*SMARCA4*	CARM1	Intra	GA II	5	Y	chr19:10955827-11033958	chr19:10843252-10894448	8	0.041	6	0.041	
MCF7	*BCAS4*	ZMYND8	Intra	GA II	4	Y	chr20:48844873-48927121	chr20:45271787-45418881	7	2.107	15	3.860	Yes
MCF7	*PVT1 (BC041065)*	MYC	Intra	GA II	4	Y	chr8:128875961-129182681	chr8:128817496-128822862	27	1.186	3	1.186	Yes
MCF7	*TRIM37*	RNFT1	Intra	GA II	3		chr17:54414781-54539048	chr17:55384504-55396899	14	3.515	2	3.412	Yes
MDA-MB-361	*TMEM104*	CRKRS	Intra	GA II	18	Y	chr17:70284216-70347517	chr17:34871265-34944326	9	2.327	7	1.529	Yes
MDA-MB-361	*TANC1*	MTMR4	Inter	GA II	12	Y	chr2:159533391-159797416	chr17:53921891-53950250	27	0.000	6	1.658	Yes
MDA-MB-361	*TOX3*	GNAO1	Intra	GA II	7		chr16:51029418-51139215	chr16:54782751-54939612	10	−0.157	19	0.281	
MDA-MB-453	*MECP2*	TMLHE	Intra	GA II	8		chrX:152948879-153016382	chrX:154375389-154495816	8	1.611	11	1.602	Yes
MDA-MB-453	*MYO15B*	MAP3K3	Intra	GA II	4		chr17:71095733-71134522	chr17:59053532-59127402	3	0.543	10	0.494	
MDA-MB-468	*UBR5*	SLC25A32	Intra	GA II	8		chr8:103334744-103493671	chr8:104480041-104496644	18	0.070	4	0.927	Yes
MDA-MB-468	*ARID1A*	MAST2	Intra	GA II	5	Y	chr1:26895108-26981188	chr1:46041871-46274383	10	0.266	23	0.818	
MDA-MB-468	*EGFR*	POLD1	Inter	GA II	5		chr7:55054218-55203822	chr19:55579404-55613083	17	4.944	4	0.732	Yes
MDA-MB-468	*RDH13*	FBXO3	Inter	GA II	3		chr19:60247503-60266397	chr11:33724866-33752647	2	0.853	3	1.507	Yes
UACC-893	*FBXL20*	CRKRS	Intra	GA II	31	Y	chr17:34662422-34811435	chr17:34871265-34944326	17	2.069	7	4.175	Yes
UACC-893	*CCDC6*	ANK3	Intra	GA II	27	Y	chr10:61218511-61336420	chr10:61458164-61570752	17	0.890	13	0.890	
UACC-893	*grb7V*	PPP1R1B	Intra	GA II	23	Y	chr17:35152031-35157064	chr17:35038278-35046404	1	4.843	2	4.843	Yes
UACC-893	*MED1*	IKZF3	Intra	GA II	9	Y	chr17:34814063-34861053	chr17:35174724-35273967	4	3.908	10	4.843	Yes
UACC-893	*EIF2AK3*	PRKD3	Intra	GA II	5		chr2:88637373-88708209	chr2:37331149-37397726	8	1.213	8	1.278	Yes
ZR-75-1	*FOXJ3*	CAMTA1	Intra	GA II	10		chr1:42414796-42573490	chr1:6767970-6854694	17	−0.380	10	−0.089	
ZR-75-1	*GPATCH3*	CAMTA1	Intra	GA II	10		chr1:27089565-27099549	chr1:6767970-6854694	3	−0.225	10	−0.089	
ZR-75-1	*C1ORF151*	RCC2	Intra	GA II	9		chr1:19796057-19828901	chr1:17605837-17637644	4	−0.013	4	−0.225	

Fusions with a recurrent partner are highlighted in yellow.

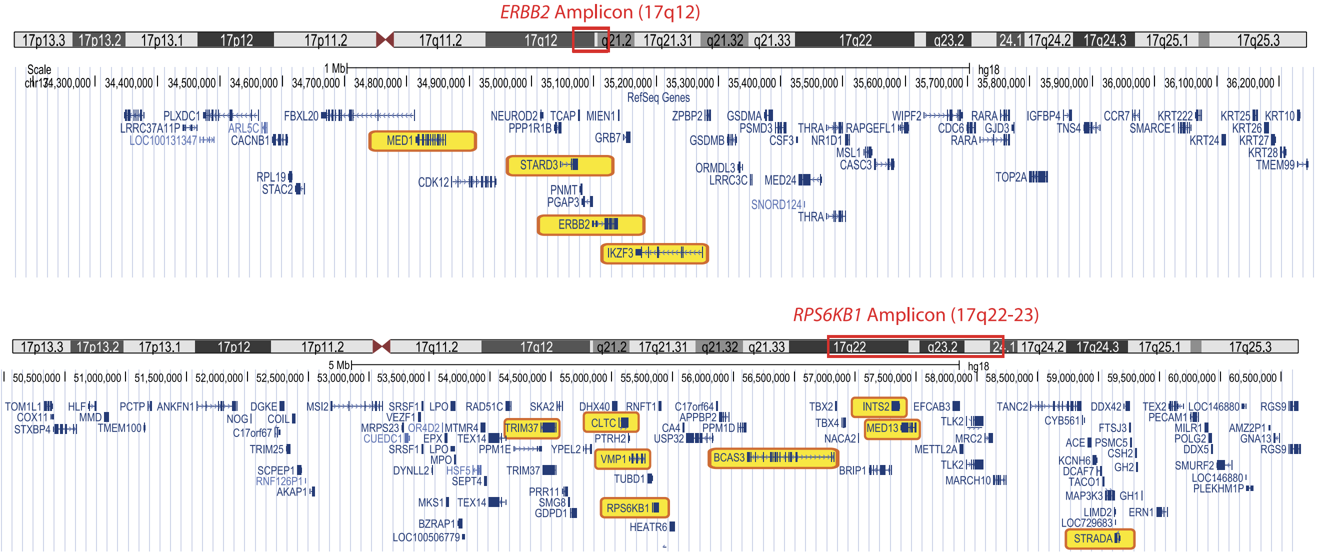
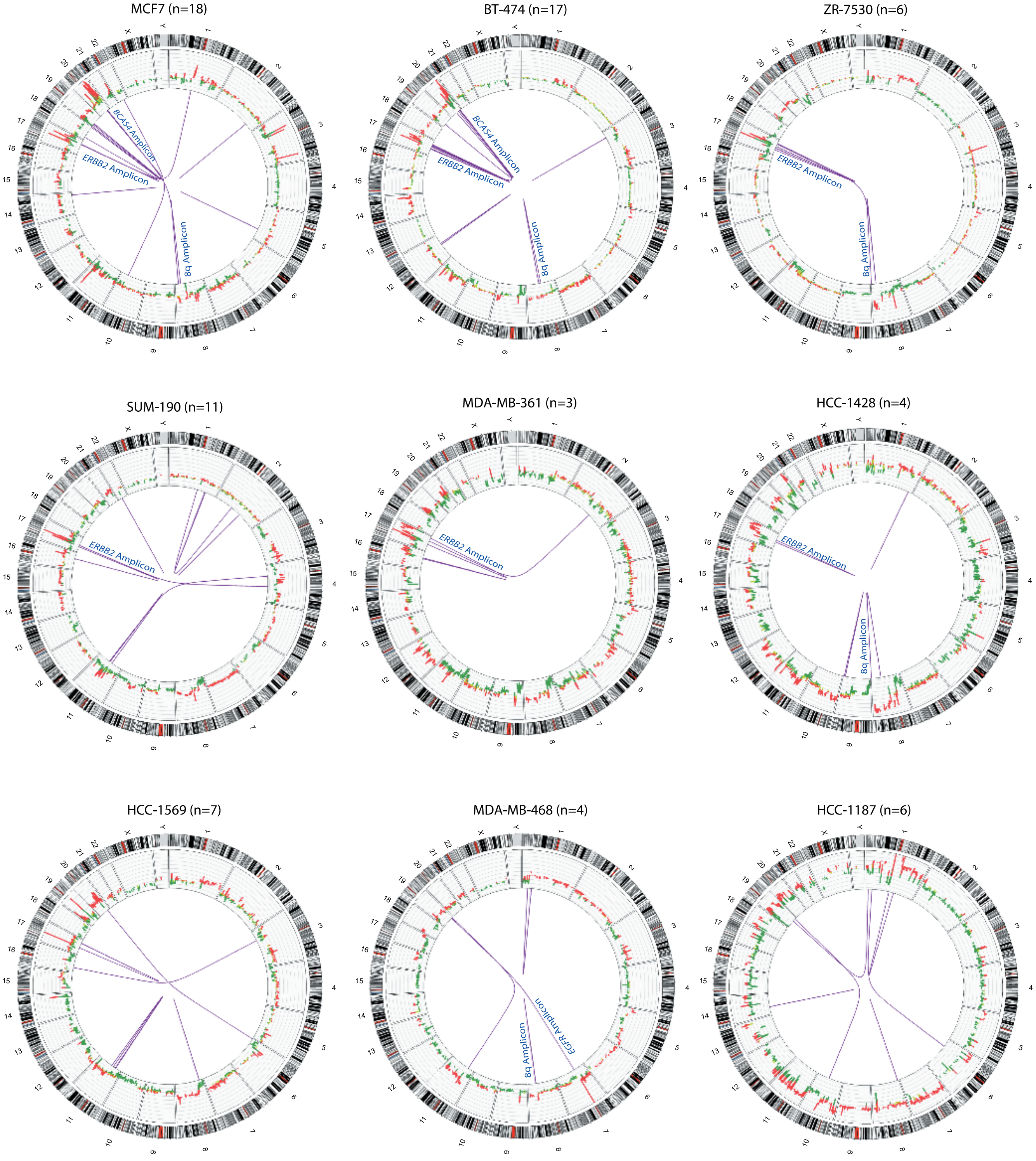


Figure W1. UCSC tracks displaying the *ERRB2* and *RPS6KB1* amplicons, with fusion genes highlighted in yellow.



MCF7 (n=18)

BT-474 (n=17)

ZR-7530 (n=6)

SUM-190 (n=11)

MDA-MB-361 (n=3)

HCC-1428 (n=4)

HCC-1569 (n=7)

MDA-MB-468 (n=4)

HCC-1187 (n=6)

Figure W2. Graphical representation of integrative analysis of gene fusions with copy number analysis. Circos plots of the genome-wide distribution of gene fusions along with status of copy number alterations. Red and green peaks represent amplifications and deletions; purple line represents the fusions associated with amplicons and nonamplicons, respectively. "*n*" refers to the total number of fusions identified.

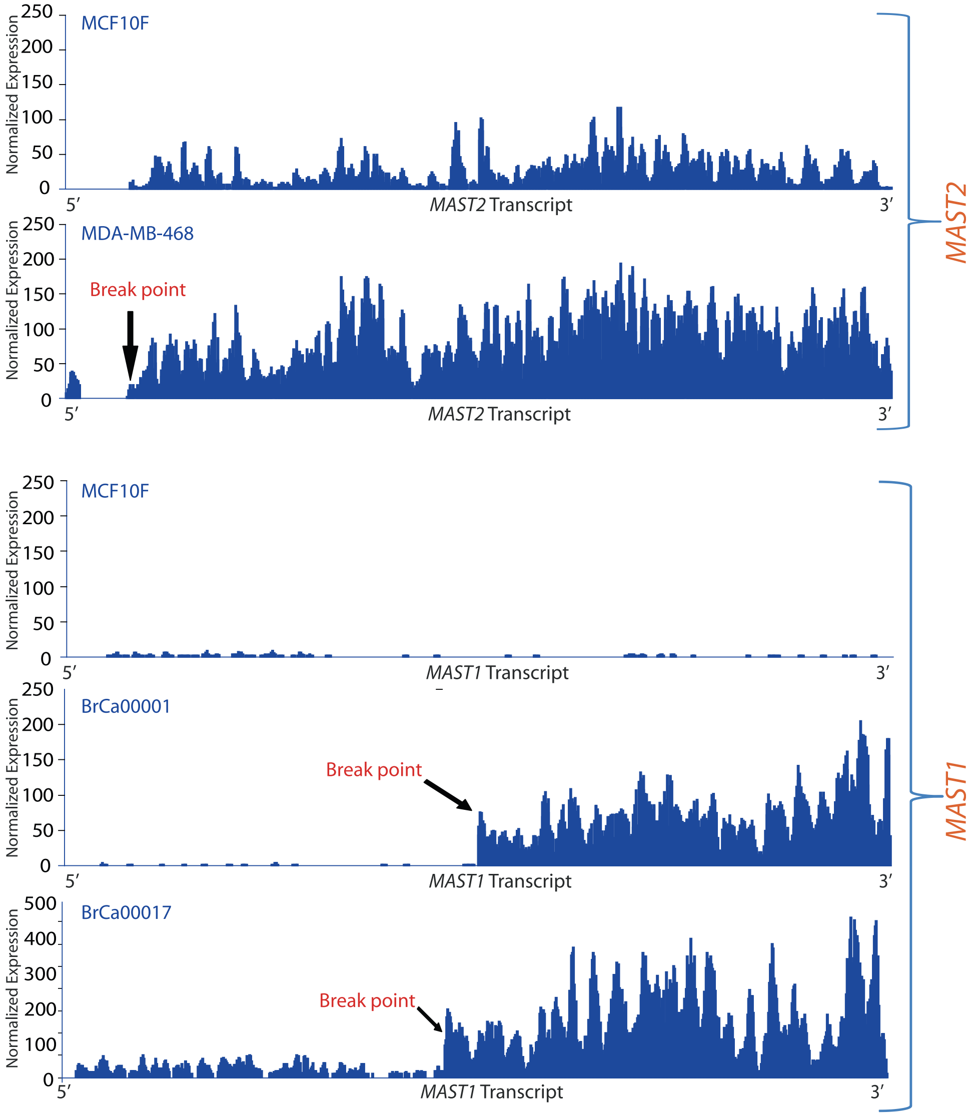


Figure W3. Plot of normalized coverage of *MAST1* and *MAST2* transcripts in *MAST* fusion-positive samples (breakpoint indicated by arrow).

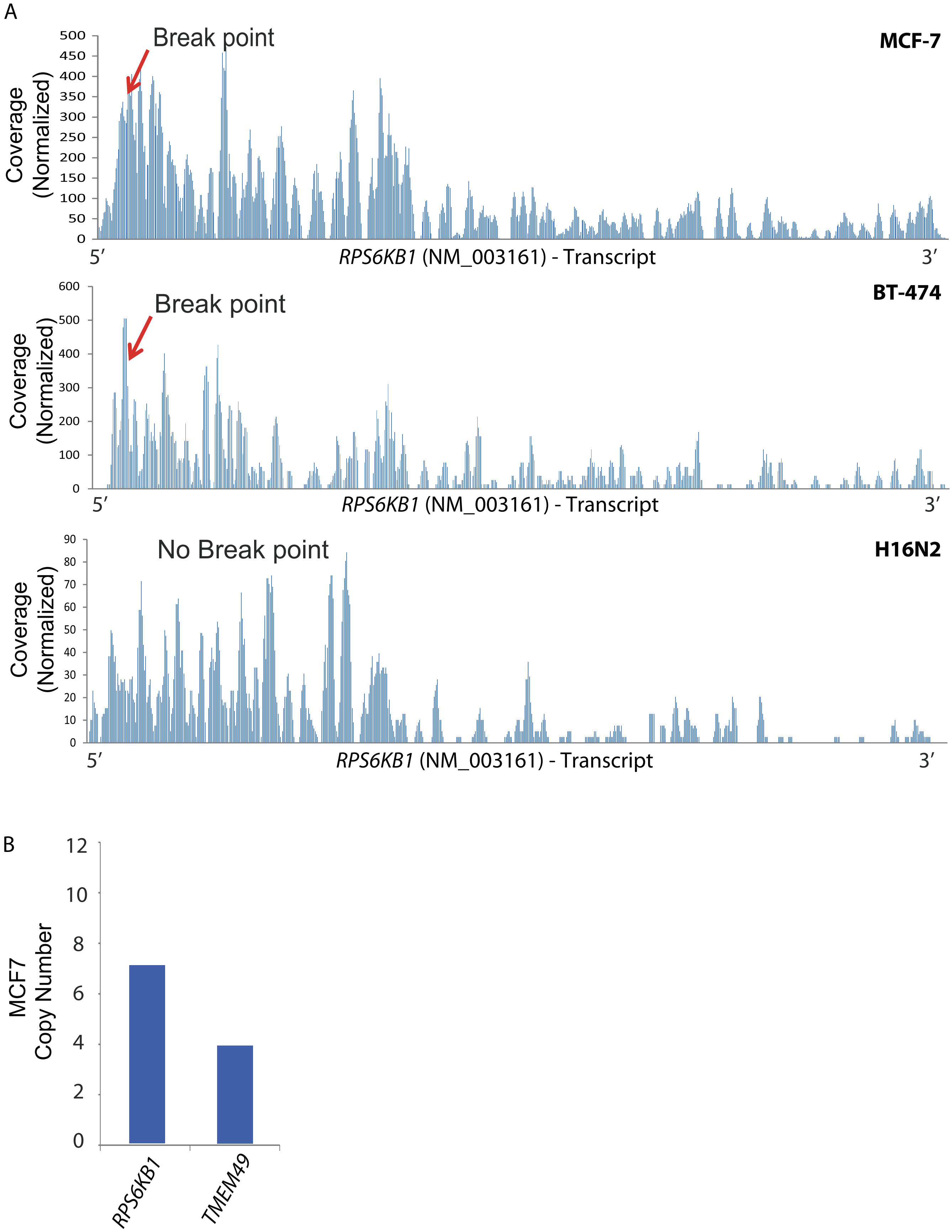


Figure W4. (A) Plot of normalized coverage of *RPS6KB1* transcript in BT-474, MCF7, and H16N2 cell lines. (B) Bar graph representing the copy number of *RPS6KB1* and *TMEM49* in MCF7.